technische universität
dortmund

Doing Wrong with Others

Multi-Agent Consequentialism as a Solution for the Collective Action Problem

by

Kevin Baum

January 8, 2025

Defended on March 22, 2024
(Revised Version for Publication)

Thesis for the degree of Doctor of Philosophy

Submitted to the Faculty of Humanities and Theology
Department of Philosophy and Political Science

First Reviewer:	Prof. Dr. Eva Schmidt	(TU Dortmund)
Second Reviewer:	Assoc. Prof. Dr. Vuko Andrić	(Linköping University)

Abstract

According to Maximizing Objective Act-Consequentialism (MOAC)—more a family of theories than a specific doctrine—the concepts of the *right* and the *best* are closely intertwined. MOAC theories assert that an action is right if and only if no alternative action has better consequences. This *criterion of rightness* seems, however, to be an expression of a more general view, according to which the 'core function' of morality consists in implicitly coordinating collective actions: those actions that, if carried out, lead to the morally best world that moral agents can collectively bring about are to be designated as right. This idea, prominently referred to as the Principle of Moral Harmony by Fred Feldman (1980), was considered unchallenged dogma within the consequentialist community until the second half of the 20th century (for example, see Baier 1958; Bentham 2007; Castañeda 1974; Mackie 1977).

However, whether MOAC theories can meet this expectation is questionable. Various circumstances—overdetermination and preemption, as well as the apparent existence of effects that, considered in isolation, are negligible but accumulate into significant harm—seem to allow the existence of collective decision situations in which combinations of actions yield collectively suboptimal results, even though no agent could have made a difference for the better by acting differently unilaterally. Consequently, such actions are apparently right according to MOAC theories, *yet* they lead to suboptimal outcomes. This puzzle, known as the Challenge of Collective Action, has questioned consequentialism for decades (see Glover and Scott-Taggart 1975; Kagan 2011; Parfit 1984; Pinkert 2015; Regan 1980; Zimmerman 1996).

This dissertation aims to reconstruct and understand the Challenge of Collective Action in its various forms and ultimately propose a novel consequentialist solution. Significantly based on game-theoretical considerations, the proposed solution results in a new and generalized MOAC theory, Multi-Agent Consequentialism, that is well-suited for multiple agents. The overarching aim of this work is to preserve the Principle of Moral Harmony as a fundamental motivation of consequentialist theorizing and, at the same time, to offer a decidedly objective-consequentialist solution to the Challenge of Collective Action.

Abstract

Gemäß *Maximizing Objective Act-Consequentialism* (MOAC) – eher eine Familie von Theorien als eine spezifische Doktrin – sind die Begriffe des *Richtigen* und des *Besten* eng miteinander verwoben. MOAC-Theorien behaupten, dass eine Handlung genau dann richtig ist, wenn keine alternative Handlung bessere Konsequenzen hat. Dieses *Kriterium der Richtigkeit* scheint indes Ausdruck eines allgemeineren Standpunktes, wonach die 'Kernfunktion' der Moral darin besteht, kollektive Handlungen implizit zu koordinieren: Es gilt diejenigen Handlungsoptionen als richtig auszuweisen, die, wenn ausgeführt, zur moralisch bestmöglichen Welt führen, die die moralischen Akteure zusammen herbeizuführen vermögen. Diese Idee, die von Fred Feldman (1980) bekanntlich als *Principle of Moral Harmony* bezeichnet wurde, galt bis zur zweiten Hälfte des 20. Jahrhunderts innerhalb der konsequentialistischen Gemeinschaft als unangefochtenes Dogma (beispielsweise Baier 1958; Bentham 2007; Castañeda 1974; Mackie 1977).

Doch ob MOAC-Theorien dieser Erwartung gerecht werden können, ist fraglich. Denn verschiedenste Umstände – Überdetermination und Präemption sowie die mutmaßliche Existenz von Effekten, die isoliert betrachtet vernachlässigbar erscheinen, sich jedoch zu großen Schäden akkumulieren können – scheinen die Existenz kollektiver Entscheidungssituationen zu erlauben, in denen Kombinationen von Handlungen kollektiv suboptimale Ergebnisse liefern, obgleich kein Akteur durch einseitiges Andershandeln einen Unterschied zum Besseren hätte bewirken können. Folglich scheinen solche Handlungen gemäß MOAC-Theorien richtig zu sein, *obwohl* sie zu suboptimalen Ergebnissen führen. Dieses Rätsel, bekannt als die *Challenge of Collective Action*, hat den Konsequentialismus seit Jahrzehnten herausgefordert (um nur einige zu nennen, siehe Glover and Scott-Taggart 1975; Kagan 2011; Parfit 1984; Pinkert 2015; Regan 1980; Zimmerman 1996).

Ziel dieser Dissertation ist es, die *Challenge of Collective Action* in seinen verschiedenen Formen zu rekonstruieren und zu verstehen, um schließlich eine neuartige, konsequentialistische Lösung vorzuschlagen. Diese signifikant auf spieltheoretischen Überlegungen basierende Lösung führt zu einer neuen und verallgemeinerten MOAC-Theorie, *Multi-Agent Consequentialism*, die für multiple Akteure geeignet ist. Das übergeordnete Ziel dieser Arbeit ist es, das *Principle of Moral Harmony* als grundlegende Motivation der konsequentialistischen Theoriebildung zu bewahren und gleichzeitig eine dezidiert objektiv-konsequentialistische Lösung für die *Challenge of Collective Action* anzubieten.

Preface

In February 2022, when President Putin initiated an illegitimate war against Ukraine, Germany found itself at a crucial juncture. Decades of increasing reliance on Russian gas had left the country vulnerable to potential supply crises. Just days after the invasion, a fervent debate erupted regarding potential shortages of this vital resource.

As the situation deteriorated sharply, appeals to citizens to take individual responsibility became ubiquitous. On social media, posts like this tweet by Rico Grimm, a German journalist, were popular: "What can you do against Putin? It's never been easier: turn down the heating, put on a sweater."[1] Echoing this sentiment, Robert Habeck, Germany's Federal Minister for Economic Affairs and Climate Action, sounded the alarm on March 30. Triggering the early warning stage for gas supplies, he reached out to the German populace with a resonant plea:[2]

> We are in a situation where I have to say clearly that every kilowatt hour of energy saved helps, and that is why I would like to combine the triggering of the early warning level for gas supplies with an appeal for help to companies and private consumers: You are helping Germany, you are helping Ukraine when you reduce your use of gas, or energy in general.

Robert Habeck's call was not to Germany as an entity but to its individual citizens. He urged them to evaluate and, if possible, adjust their behaviors in light of the looming gas shortage and its associated repercussions. The underlying message was clear: every little individual contribution, no matter how small, helps.

However, skepticism quickly followed. Critics questioned the feasibility and sensibility of a 'freezing for peace' approach. They labeled such an appeal as arrogant and cynical.[3] They pointed to the vast capacity for gas storage, the adaptability of markets, and the minuscule proportion of private consumption when juxtaposed against industrial demands. The common argument was this: individual efforts were like a drop in the ocean. So, while most agreed

[1] https://twitter.com/gri_mm/status/1499368527844757516, my translation.

[2] Cf. https://www.dw.com/en/german-economy-minister-raises-warning-level-for-gas-supplies/a-61300264

[3] Cf. https://www.t-online.de/finanzen/unternehmen-verbraucher/id_91797770/frieren-fuer-frieden-das-zeugt-von-purer-arroganz.html.

that it was indeed crucial for Germany and the broader EU to cut down on consumption and move away from dependency on Russian gas, they viewed individual efforts as comparably inconsequential.[4] How can individuals be expected to shoulder the burden of past political miscalculations—especially given that the consequences of each individual's behavior seem so negligible?

This thesis, essentially, traces the question of whether the skeptics' stance is valid. If they are correct, it means that although *we*—considered here as a rather loose collective of rather uncoordinated individual citizens—are morally required (in some sense) to take *collective action* (like reducing gas consumption), each *individual* might justifiably abstain because their isolated efforts will (probably) not add any significant contribution. Thus, the question driving this project is:

> (Q) If each of us always does what is morally right, are we guaranteed to bring about the morally best results that we could together bring about?

For the past decade, this question has occupied my mind. I was particularly attracted to (Q) because, although it is a theoretical question, it underlies many practical, real-life challenges, some of which are among the biggest of our time. For instance, the man-made climate crisis can be discussed morally against the background of (Q) on several levels (cf. Budolfson, McPherson, and Plunkett 2021; Gardiner 2011; Sinnott-Armstrong and Howarth 2005), as can questions of a moral obligation to resist oppressive social systems or regarding a duty to vaccination. Thus, while this inquiry primarily probes a theoretical moral-philosophical dilemma, its implications are far from theoretical. In fact, they are of as much practical importance as moral questions can be.

This thesis addresses (Q) as a theoretical challenge, though. It does so from the perspective of a specific notorious family of moral theories for which that challenge is arguably particularly pressing, namely Maximizing Objective Act-Consequentialism (or simply MOAC for short). By the end of my investigation, I will have shown that, contrary to popular belief, an advanced MOAC theory can affirm (Q) for a broad class of relevant cases. However, I will also argue that (Q) must be denied for another class of critical but particularly relevant cases, even though this is (somewhat surprising in the light of the prevailing debate) not a deal-breaker for MOAC.

[4] This argument is replicated in the March first episode of the podcast *Deutschlandfunk: Der Tag* (https://www.deutschlandfunk.de/01-03-2022-der-tag-blau-gelb-im-europaparlament-dlf-21205766-100.html) and the *Die Zeit* article "Energiesparen gegen Putin" on March 13 (https://www.zeit.de/wirtschaft/2022-03/gasversorgung-russland-ukraine-krieg-embargo-energiesparen).

On Style and Methodology

I am convinced that scientific progress is possible in the humanities, including philosophy. But it comes with the necessity to wrestle with thoughts, words, and concepts not only until what is to be said is said—and no more and no less—but also until what was complicated to understand is easy to understand. In this regard, I subscribe to the ideals of analytic philosophy. Precision in thought and language, structure and systematicity, arguments and conceptual analysis, and a readiness to apply formalism where useful are essential features of doing good philosophy.

Meeting these demands is not easy. The path to initial knowledge is winding and is never the shortest and straightest. However, once you have reached your destination, a better and more convenient path can typically be found—provided you make the effort. The task of the scientific philosopher is to show others such a better and more convenient path, or at least a suitable approximation, possibly taking into account one or the other panorama or sight worth seeing that one found on one's own way to the top. It is undoubtedly not good philosophy to force the readers to take the the complicated and stressful original path.

While I have been struggling to find such a convenient path, I think I have found an acceptable one. As is so often the case, however, one is never fully satisfied at the end of such an endeavor. For sure, you can always find another, better path. Where I have failed, I hope to have failed at least in an engaging and instructive way.

The purpose of this book is to strengthen MOAC, one of the most significant and influential families of moral theories, by addressing one of its most pressing internal theoretical challenges. Achieving this requires tapping into the (often implicitly shared) conceptual toolbox of MOAC and expanding it where needed. It also demands a thorough understanding of certain claims and criticisms so that they can be refuted effectively. Formal methods can prove indispensable here.

Part of my task, therefore, has been to "go formal" whenever it promises greater precision, clarity, and insight, and to avoid unnecessary formalism wherever it might obscure understanding. Striking this balance has been a delicate endeavor, and while I may have faltered at times, I hope I have succeeded more often than not.

However, this could have easily become another project. Many of this book's core ideas and claims could certainly be proved and derived within a *thorough* formal framework. But that would then show that the ideas I hold and the points I make are valid *within* that specific formal system. Until it is shown that the relevant system captures what we (or at least the champions of camp MOAC) call morality, however, this would not show what I want to

show. Such a project would be one of deontic or, more precisely, multi-modal logic. At times, my project was on the verge of becoming this project, which would undoubtedly be exciting in its own right. But it is neither the project I started with nor the one I ended with. It is not this project.

In this work, as a consequence, I employ what can best be described as a kind of semi-formalism. While resonating with known frameworks, the notations, notions, concepts, and structure used in this work are distinct from existing ones and specially tailored for my purposes. If I were to align it more closely with any established framework, stit semantics would be the chosen one (where "stit" stands for "seeing to it that", cf. Belnap, Perloff, and Xu 2001; Horty 2001; Horty and Belnap 1995). Stit semantics offers a structured platform for multi-modal logics of action, drawing inspiration from Arthur Prior's branching time model (Prior 1955, 1967). In many ways, the formalism presented in this thesis can be perceived as stit 'undercover', enabling the utilization of stit's core concepts without being constrained by its rigorous formalities. I owe a profound debt of gratitude to John Horty's seminal works, especially his in-depth explorations in the field (cf. Horty 2001, 2019).

Acknowledgements

A project of this scale, developed over so many years, cannot really have just one creator. It takes a whole community of supportive friends and colleagues—a village of helping hands—to see it through.

First and foremost, I would like to thank the people who encouraged me and gave me the courage to continue whenever I felt like giving up. My deepest gratitude goes to my wife, Deborah, who has unfailingly stood by me since before I even started this project; I owe her more than I can ever repay. Next is Sarah Sterz, who was first my student, then my research assistant, then a colleague, and along the way became a close friend. I am also indebted for 'spiritual support' to Thorsten Helfer, Marija Slavkovik, Holger Hermanns, Markus Langer, Stephan Padel, Timo Speith, Susanne Mantel, Stephan Schweitzer, Jonas Harney, Robert Reick, and many others. Without all of you, this book simply would never have come to be.

Professionally, I owe a great deal to many of the individuals named above, but especially to my long-time supervisor, Christoph Fehige, and my friend and colleague Oliver Petersen. All participants of the *Doctoral Colloquium for Practical Philosophy* (which now goes by a different name, though I will forever call it "the DKPP") have consistently provided valuable feedback; they had to endure my all-too-frequent half-formed and unfinished ideas concerning collective decision-making. Special thanks are due to Jeff Horty, who faced my constant questions and musings at several seminars at Schloss Dagstuhl—always among the scientific and intellectual highlights of these years. What

bad luck for him that I appeared at those seminars, but what tremendous luck for me in so many ways. Everyone involved in the *Explainable Intelligent Systems* project and at the *Center for Perspicuous Systems*—thank you for listening (or patiently enduring my complaints) about this endeavor. You helped me more than you might realize. Finally, this work would not reach its current level of linguistic quality without the efforts of Christopher von Bülow and Laura Stenzel. Thank you both!

Many people over the years have believed in me strongly enough to fund my work and grant me the freedom to develop and unfold. In addition to the two professors already mentioned, Christoph Fehige and Holger Hermanns, this is especially true of Ulla Wessels and my current boss, Verena Wolf. Thank you all; I hope I have never blatantly disappointed you. Likewise, my thanks go to Eva Schmidt and Vuko Andrić. You not only provided invaluable feedback but were also there when I had to make significant changes to my Ph.D. plans in this final year. I am *so* indebted to you that I can only hope you never decide to redeem that debt.

Last but not least, I must mention another special person—a friend whom I dearly miss. Daniel Oster was a wonderful colleague who rescued me more than once when I was stuck, whether in terms of content or emotionally. Damn, Daniel, I miss you.

Contents

Chapter 1

Introduction

The relationship between the morality of actions and the moral quality of their consequences is a topic of much debate. A significant portion of the philosophical community even believes that it is the primary function of morality to guide us toward morally optimal outcomes. Thus, they typically find themselves aligned with the following principle:

(R) If someone does what is morally right, they are guaranteed to produce the morally best possible results that they could bring about.

The question at the core of the present project is whether this intricate relationship translates to *collective* contexts. This boils down to the following question:

(Q) If each of us always does what is morally right, are we guaranteed to bring about the morally best results that we could together bring about?

Those who believe in (R) will arguably be inclined to affirm (Q). Of course, other conceptualizations of morality do *not* emphasize the relationship between the right and the best so much. Accordingly, many proponents of those conceptualizations will readily reject (Q) without hesitation. However, this project is tailored to (R) believers.

It has been argued for decades that, for a particular, highly influential family of moral theories, as famous as they are notorious, which I will hereafter call Maximizing Objective Act-Consequentialism (or "MOAC" for short), denying (Q) is no viable option. *Act-Consequentialist*[5] theories of morality

[5] In the following, I will mostly drop the prefix "act-". There are other kinds of consequentialist theories, for example, those that operate on the level of motives (Sverdlik 2011) or rules (Brandt 1959; Hooker 2023). Nevertheless, these theories will not play a role in my project, so whenever I write of "consequentialist theories", I shall always mean *act*-consequentialist theories.

make the moral quality of the consequences of actions (and their alternatives) the *sole* measure of the morality of those actions. *Objective* consequentialist theories limit their theoretical considerations to the moral qualities of the *objective* consequences of actions, i.e., to those consequences that would be the case if a specific alternative was performed. Which consequences are to be expected or are actually expected by some agent is thus irrelevant to objective consequentialist theories. *Maximizing* consequentialist theories ask us to *maximize* the moral qualities and not, for instance, to produce good or 'just' sufficiently good consequences (in the sense of suboptimal consequences being good enough, cf. Slote and Pettit 1984).

As the name implies, MOAC theories are that subclass of consequentialist theories that are objective and maximizing. They are still a *family* of theories—rather than a single, specific theory—in that there is disagreement among MOAC theories about the nature and specifics of the moral quality of consequences. It follows that these theories are characterized by a common criterion of *rightness*,[6] the Maximizing Objective Criterion of Rightness (or, more briefly, simply "MOCOR"), which we can tentatively express as follows and obviously corresponds to (R) above:

Criterion 1.1 (MOCOR – tentative) *An action is right if and only if there is no alternative action that would lead to better consequences.*

This brings us back to the initial question (Q): On the one hand, MOAC theories seem committed to answering the question in the *affirmative* because they put a consonance between right action and morally optimal outcomes at the center of their philosophical theorizing. On the other hand, certain collective decision scenarios present a conundrum in this regard because, in some situations, the consequences of our actions are inextricably linked to the decisions of others, and this interdependence can often be mutual for all involved. In such cases, all agents can act such that they together bring about suboptimal, even disastrous consequences, such that apparently none of them could have changed anything for the better by acting differently. In such cases, however, MOCOR seems committed to assessing that everyone acts rightly, implying that MOAC theories have to answer (Q) in the negative.

[6] This work is limited to questions of moral rightness and operates mainly on the widely shared assumption that morally right actions are precisely those actions that are morally permissible (although more needs to be, and will be, said about the precise relationship between these different kinds of moral status). Further, from this point on, I mean by "rightness" by default *moral* rightness. The focus on rightness also largely corresponds to language in much of the literature on the CHALLENGE relevant to this project. In some quotations, however, we will encounter talk of an agent who *ought* to act in specific ways. In these cases, such talk typically can and should be understood in terms of rightness by adopting that an agent ought to perform one of the alternatives right for them and, thus, that if there is only one right (i.e., permissible) option, the agent ought to perform that alternative.

I call this and similar challenges we will encounter later the Challenge of Collective Action (or simply "the CHALLENGE" for short). Solving it within the framework of MOAC theories is the primary goal of this work.

The CHALLENGE has occupied moral philosophers for quite some time. At least since the 1970s (cf. Glover and Scott-Taggart 1975) and in particular in the aftermath of Donald Regan's book *Utilitarianism and Co-operation* (1980), consequentialists have endeavored to save MOAC in the face of the CHALLENGE. Some have taken it as an occasion to jettison fundamental consequentialist beliefs (Feldman 1980; Jackson 1987; Parfit 1988; Sinnott-Armstrong 2005) regarding the (guaranteed) consonance between right action and optimal results. Others have used it as a reason for substantial modifications of the consequentialist criterion of rightness (Parfit 1984; Zimmerman 1996) or even for abandoning (act-)consequentialist grounds altogether (Regan 1980). Over the years, the debate has spread to subjective consequentialist terrain (Budolfson 2019; Hedden 2020; Kagan 2011; Pinkert 2015; Portmore 2018). Some authors have discussed the CHALLENGE from other, more general perspectives (Andreou 2014; Nefsky 2011) and asked how far it extends into non-consequentialist territory (Killoren and Bekka Williams 2013). This is just a sample of the vast and rich literature on the CHALLENGE.

The presentation so far may have given the impression that the CHALLENGE is mainly theoretical. Nevertheless, the CHALLENGE lies at the heart of some of the most pressing practical issues of our time. For instance, think of the anthropogenic climate crisis. Although it is not certain what their concrete form will be, the consequences of our collective greenhouse gas emissions will undeniably be catastrophic. However, it is by no means obvious what this means for the moral status of all the myriad of actions of the individual inhabitants of Earth, which, in sum, are causative for at least a significant part of those emissions. After all, the emissions of a single, easily avoidable short-distance car trip (or even all the consequences of one individual's consumption and mobility decisions over time) are 'globally' negligible. As Walter Sinnott-Armstrong once put it: "global warming and climate change occur on such a massive scale that my individual driving makes no difference to the welfare of anyone" (cf. Sinnott-Armstrong and Howarth 2005).

At the same time, however, avoiding *prima facie* contributing actions is often accompanied by morally significant individual costs. Thus, given the apparent individual inefficacy, it seems that while no individual effort to save greenhouse gases will make a morally relevant *positive* difference, many such decisions will make morally significant *negative* ones. This finding is entirely independent of whether, at the end of the day, sufficient greenhouse gas emissions are avoided overall. Therefore, the question arises of how a moral demand for personal sacrifice can be (morally) justified. If a sufficient

number of other agents 'do their share', my individual sacrifice is a waste of moral value; and if the collective effort falls short, my individual saving does not change a thing for the better. Thus, it seems morally right (at least in light of MOCOR) to *not* restrict myself, no matter what the others do. Consequently, it seems better overall if I took that easily avoidable but comfortable car trip. In other words, individual losses appear to trump collective gains. But if all this is true, there are situations in which acting rightly leads to suboptimal and even catastrophic consequences—which brings us back to (Q) and the CHALLENGE. Therefore, although this project deals with the CHALLENGE in a theoretical setting and rather abstract terms, it also concerns some of the most pressing practical issues of our time, at least from the point of view of 'camp MOAC'.

This project attempts to find a solution to the CHALLENGE that remains true to consequentialist core tenets. Thus, this thesis is a consequentialist project, i.e., limited to a consequentialist perspective. At the same time, the contribution of this work to the advancement of consequentialist theorizing is by no means merely to dispel the CHALLENGE, which is just one particular, albeit perilous and significant challenge for MOAC. For, as will become apparent in the course of this thesis, the CHALLENGE is only a symptom of a more profound failure of objective act-consequentialist theories. Principled and fundamental difficulties arise for instances of MOAC from the fact that the consequences of actions may well depend on the actions of other agents, and the CHALLENGE arises only from the inadequate methods that consequentialists chose for their approach to collective decision situations in the past. Therefore, in addition to the specific goal of mastering the CHALLENGE, this work has a more general and theoretically more fundamental goal: it is about nothing less than the search for an objective consequentialist moral theory that can do justice to the fact that each of us is only one among many. Accordingly, this book makes it its mission to develop a *consequentialist multi-agent moral theory*.

1.1 The CHALLENGE in a Nutshell

At its core, this project is an attempt to defend Maximizing Objective Act-Consequentialism, and thus a very particular family of moral theories, against a specific challenge: the CHALLENGE. However, there is no canonical representation of the CHALLENGE, but rather several different formulations. These differ, on the one hand, in terms of their potential impact and, on the other hand, in terms of the theoretical conditions necessary for their formulation. In the best tradition of analytic philosophy, this work tries to defend MOAC against the *strongest* version of the CHALLENGE for MOAC. However, as it turns out, the strongest version of the CHALLENGE is *not one* variant of

it. Instead, it consists of several hierarchically ordered variants that back each other up. The first line of attack consists of a version which I call Internal Challenge (where "internal" is used to indicate that the challenge is theory-internal) that operates with a fair amount of (rather convincing) presuppositions. Its success would invalidate moac, establishing the inadequacy of all moac theories. If it fails, it can be replaced by other variants, most notably the so-called No-Difference Challenge and, as we will see, also with a pre-theoretic version as a second fallback (that I shall call the Trilemma). These variants get by with significantly weaker presuppositions, but, if successful, would also have a less dire impact. Accordingly, to rebut Internal Challenge is the central goal of this thesis. In the following, this variant will be presented in rough sketches, whereby the basic intuitions invoked in the previous section shall get some additional theoretical underpinning. The other two versions will have to wait until Chapter 3, a deep dive into the state of the debate that follows a preparatory chapter (Chapter 2).

Internal Challenge is based on the insight that, arguably, mocor expresses (or is at least motivated by) a more general view on morality, viz.:

View 1.1 (Congruence) *The right and the best are congruent in the sense that doing what is right goes (necessarily) hand in hand with bringing about the morally best consequences that can be brought about.*

It seems not to be far-fetched that Congruence is the fundamental assumption underlying mocor. In this respect, mocor does not just arise out of thin air but is rather to be understood as one possible explication of Congruence in the form of a criterion of rightness. mocor would, in this case, be intended to capture the spirit of Congruence and, thus, should be respected to live up to it.

Congruence and mocor not only go well together but are closely related. That performing one of the actions with the best consequences comes with the best consequences that can be brought about seems to be a truism. However, it is also tempting to read Congruence in a *collective* sense, as an expectation that morality has the general property of 'marking the path to moral optimality'. In other words, according to such a reading, an appropriate moral theory comes with a criterion of rightness that picks out exactly those actions that, if consistently performed, necessarily lead to the best possible outcomes that these agents can bring about. Several authors have proposed formulations of this broader claim, which is rather a second-order claim concerning the nature of morality itself: Donald Regan (1980), Fred Feldman (1980), who tracked the idea down through history, and, more recently, Felix Pinkert (2015) and Douglas Portmore (2018). In a first approximation based on their work, we might capture the idea like this:

Principle 1.1 (Collectively Maximizing – tentative)
If all agents act rightly, then they are guaranteed to produce the morally best outcome they could together bring about.

MOCOR and Collectively Maximizing certainly seem like a perfect match at first sight. It might even be tempting to think that a rightness predicate explicated by MOCOR virtually *guarantees* the truth of Collectively Maximizing. Indeed, one might conclude that if all the agents of some collective perform one of the actions with the morally best possible consequences, this *must* lead to the morally best possible consequences that this collective can bring about together. In coming to such an (as it turns out) hasty conclusion, one would be in the best company. Jeremy Bentham, for instance, one of the founding fathers of utilitarianism and arguably one of the most famous exponents of MOAC (cf. Woodard 2019), writes confidently that his *Principle of Utility*, "which approves or disapproves of every action whatsoever, according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question" (Bentham 2007), is "capable of being consistently pursued; and it is but tautology to say that the more consistently it is pursued, the better it must ever be for humankind" (ibid.).

Based on this seemingly obvious insight, various authors, including several consequentialists, have elevated Collectively Maximizing to criterion of adequacy for moral theories. Following Fred Feldman's extensive and thorough work on this 'collective reading' of Congruence (cf. Feldman 1980), we might call that expectation the Principle of Moral Harmony (or shorter just "PMH"). This name has prevailed in the literature to this day (cf. Portmore 2018). The principle is meant to express the idea that morality 'lights the way' to moral optimality not only on an individual but also on a collective level, and that, thus, we can make Collectively Maximizing a requirement for any potentially adequate moral theory. We can capture PMH[7] tentatively in terms of a simple necessary condition:

Criterion 1.2 (Moral Harmony (MH) – tentative) *A moral theory is adequate only if it is true that if all agents act rightly (according to this theory), then they are guaranteed to produce the morally best outcome they could bring about together.*

Logically equivalent,[8] but slightly reformulated (especially concerning the contraposition applied to the consequence), the statement can also be expressed negatively:

[7] I'll use "PMH" for the general idea and "MH" for its various formulations.

[8] The only reformulation here that is not purely syntactic but semantic/conceptual is that I suggest "at least one of the agents acted wrongly" where, strictly speaking, only the weaker "at least one of the agents acted not rightly" is logically warranted. The relevance of this difference will be explored in more detail later.

Criterion 1.3 (Moral Harmony (MH) – tentative, contraposition)
If a moral theory is adequate, then, if the agents in a collective decision situation produce a morally suboptimal outcome, (necessarily) at least one of the agents acted wrongly (according to this theory).

Obviously, this takes us back to the question at the beginning of this introduction, which was

(Q) If each of us always does what is morally right, are we guaranteed to bring about the morally best results that we could together bring about?

Obviously, if Moral Harmony is correct, then every adequate moral theory will necessarily answer (Q) in the affirmative.

If Bentham is right, Congruence, MOCOR, and Collectively Maximizing (or, respectively, MH) fit together perfectly. In particular, Collectively Maximizing would simply be an *implication* of MOCOR, which means that MOCOR satisfies the adequacy criterion trivially. However, as tempting as it is to elevate an apparent consequence of one's criterion of rightness to a criterion of adequacy for moral theories, this is dangerous when that very adequacy criterion backfires. After all, as is so often the case in philosophy, seemingly obvious, trivial, and uncontroversial statements prove to be major challenges.

In this case, it is a simple observation that threatens the foundation of consequentialist theorizing: each of us is just one among many and, moreover, none of us lives in a neatly isolated bubble.[9] We all interact and stand in many kinds of interdependent relationships. Specifically, the results of what any one of us does are usually codetermined by what the others do. Hence, what one person can achieve (or screw up) usually also depends on what others do. However, if the results of our actions depend, at least sometimes and partly, on what others do or will do, then it does not seem settled at the time of these actions what their total outcomes will be. Some consequences will conditionally depend on what others do.

Precisely these features of underdetermination and dependencies, both to be examined in more detail later, are what threaten MOAC's fulfillment of MH—and thus pose an existential threat to consequentialist theories. For this form of mutual interdependencies can lead to unfortunate situations in which particular combinations of actions lead to morally suboptimal or even disastrous results, while it seems that no individual agent's change in behavior would have made any difference for the better. It proves helpful to have a name for such cases that contain such combinations of actions:

[9] Or, more eloquently put by John Donne (1923): "No man is an island, Entire of itself. Each is a piece of the continent, A part of the main."

Definition 1.1 (Troublemakers – tentative) *A collective decision situation is a Troublemaker if and only if there is a troublesome combination of options therein, i.e., the agents can act in ways such that*

> **(Collective Suboptimality)** *together they would produce a morally suboptimal outcome and*
>
> **(Individual Optimality)** *none of them could make a difference for the morally better by unilaterally acting differently.*

It must be pointed out that Collective Suboptimality is in no way meant to refer to a collective action in the sense of a joint, coordinated action. That the agents "together produce" a suboptimal outcome through their actions simply states that the mere combination of the individual actions results in precisely these consequences, completely independent of shared intentions, joint decision-making procedures, or implicit coordination with each other. In fact, the absence of such a group-agent-like character generally makes Troublemakers such trouble. This is because, due to the absence of a possibility to coordinate, that path out of the Challenge is blocked.

Troublemakers raise the Challenge, as they apparently reveal an inconsistency between MOCOR and Collectively Maximizing: Assume that some agents find themselves in a Troublemaker scenario and that the agents perform a troublesome combination. According to Collective Suboptimality, they then produce suboptimal outcomes. Thus, according to Collectively Maximizing, at least one of them must have done wrong. However, given Individual Optimality, MOCOR seemingly entails that all of them did right. Thus, the Challenge reveals a serious tension between MOCOR and Collectively Maximizing, two principles that are meant to explicate the same basic conviction, viz. Congruence. The Challenge (as Internal Challenge) then is apparently this:

> If
>
> - MOAC theories are characterized by being theories embracing MOCOR,
> - and if the champions of MOAC are committed to accepting MH,
> - and if the existence of Troublemakers proves that MOCOR violates Collectively Maximizing,
>
> then *MOAC theories are inadequate according to their very own standards.*

In this sense, then, as Shelly Kagan has put it, MOAC seems "to fail even by its own lights" (Kagan 2011, p. 108). Here is a tentative proposal for the Challenge in this theory-specific sense (the Internal Challenge) in terms of a tabular argument:

The ARGUMENT – tentative

$P_{\exists\text{TROUBLE}}$: There are Troublemakers: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

P_{MOCOR}: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

P_{MH}: If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (according to this theory).

$C_{\neg\text{ADEQ}}$: MOAC is not an adequate moral theory.

This variant of the CHALLENGE is much more than a philosophical puzzle for consequentialists. It is meant to reveal a truly serious theory-internal inconsistency, a matter of life or death for one of the most influential and time-honored families of moral theories.

This book aims to show that the situation is not as hopeless as it might seem at this point. I claim that, contrary to the prevailing consensus, MOAC *has* the conceptual space to address the CHALLENGE without grave, origin-denying modifications other approaches suggested, and without having to abandon important theoretical or motivating ground. The goal of my project is to show that, somewhat surprisingly, MOAC can, without betraying its origins and core motivation, master the CHALLENGE.

1.2 Structure

This thesis has two parts, which have different objectives. The first part is reconstructive and preparatory in nature, while the second part is constructive and serves to develop and defend my own approach.

The reconstructive Part I consists of three chapters. Chapter 2 is preparatory and consists of notes on the nature and structure of individual decision situations and consequentialist theorizing, including useful notions and formalisms. Then, in Chapter 3, I sort through various strands of the debate on the CHALLENGE. In doing so, I differentiate between three different variants

of the Challenge (the Internal Challenge, the No-Difference Challenge, and the Trilemma) and explain how they relate to each other. In particular, I argue that 'the' strongest version of the Challenge is the Pyramid, a three-layered amalgam of these variants. Along the way, I distinguish different kinds of Troublemakers and collect some examples that will accompany us throughout this project. I conclude part I with Chapter 4, where I discuss what satisfactory approaches to Challenge would have to look like. I do this structurally by characterizing so-called solution spaces, which describe the space of possible 'theoretical moves' that can lead out of the Challenge. But I also do it substantially by introducing, collecting, and justifying a set of criteria. In this context, I also explain why it is justified to exclude a specific type of Troublemakers, which I call Cumulative Effects Cases, from the scope of my project. Afterward, I will briefly discuss a rough taxonomy of existing approaches, discuss three carefully selected approaches from the literature and give some reasons why they are not satisfactory. This suffices to motivate my own approach.

The constructive Part II comprises four chapters. In Chapter 5, I introduce some further preliminaries regarding collective decision situations and the assessments that moac theories give for them. Next, in chapter Chapter 6, I scrutinize the common understanding of the Internal Challenge as sketched above. While it can be established that the Argument is not even valid, this comes at the cost of accepting that the Challenge is rather a symptom of a deeper, conceptual issue. As a result, I present what I call the Real Challenge. Mastering the Real Challenge in a way that does not reproduce the Challenge is the central goal of the remaining chapters. In chapter Chapter 7, I advocate for a fresh approach involving a conceptual framework that introduces a new kind of consequences of actions to the objective consequentialist workbench. Combined with one of various *prima facie* promising so-called *collective amendments*, moac can solve the Real Challenge. In Chapter 8, I discuss how camp moac should choose between various such amendments in a way that allows solving the Real Challenge without violating pmh. Finally, I arrive at a concrete recommendation, a generalization of moac that I call Multi-Agent Consequentialism ("mac", in short).

In the final Chapter 9, I summarize my work and dare a brief outlook on possible further work. I conclude by noting that moac's ability to adequately handle collective contexts, hitherto a major Achilles heel of camp moac, have now become a strength, a tangible competitive advantage in the constant battle for the best, most advanced moral theory.

Part I

The Challenge (and How *Not* to Solve It)

Overview of Part 1

Historically, consequentialism has been confronted by a plethora of philosophical challenges. What I call the Challenge is a crucial one—an issue that has persisted in academic discourses and catalyzed significant debates. The present dissertation positions itself within this ongoing discourse, aspiring to bolster the consequentialist framework against the Challenge and related challenges that have their origin in collective contexts. To this end, it is immensely important to first truly understand the Challenge in all its blurriness, and to unearth it in its various forms from the depths of philosophical discourse. Of course, this also includes taking a critical look at existing attempts to refute the Challenge. Only what one can talk about properly, one can hope to solve; and only those who have screened the giants can choose the shoulders on which they want to stand.

Establishing such foundations is the goal of the first part of my project. It paves the way for the subsequent project of this thesis by setting the stage for a rigorous reinterpretation of the Challenge and, finally, a robust and sustainable solution to the Challenge. Accordingly, Part I comprises three chapters. In Chapter 2, I lay some groundwork. The main goal is to start with a proper understanding of *individual* decision situations and a basic understanding of consequentialist theories (and their defining components) so that one can then move on to *collective* decision situations. Next, in the voluminous Chapter 3, the Challenge is dissected by means of the reconstruction of several variants that can be extracted from different strands in the existing literature. The culmination of this analysis is the creation of a three-layered amalgam of different variants of the Challenge, which I term the Pyramid, and which is designated as the ultimate focal point of my project. Subsequently, in Chapter 4 adequacy criteria are established, which serve as benchmarks for successful refutations of the Challenge. Additionally, that chapter explores the shortcomings of at least three prominent approaches. The second part of the thesis will then be committed to constructing my proposed resolution.

Chapter 2

Preliminaries I

In this section, I will introduce a set of concepts, notions, abbreviations, and naming conventions to ensure clarity and conciseness. Moreover, I will construct a semi-formal model that encapsulates (individual) decision situations as pertinent to (especially consequentialist) moral theories. This foundation will be instrumental in effectively communicating the vital properties and concepts that will be explored in the chapters to follow. While most of the initial portion of the thesis does not require an understanding of these preliminaries, they become increasingly important as the discussion progresses.

2.1 Individual Decision Situations

Fortunately, for a sufficient degree of precision, we do not need something as complex as a complete semantic for advanced deontic logic. The following relatively simple framework, constructed from a few key building blocks, does suffice:

Agents, Decision Situations, Options: Every now and then, an *agent*[10] finds herself in a *decision situation*, i.e., a situation where she has to choose between several *options*.

Actions: Each option has a *corresponding* action.[11] Options can be *instantiated* by *performing* the action corresponding to that option. Under

[10]By default, we may think of agents as natural persons, but there might also be group agents (List and Pettit 2011). In the context of this work, the important distinction is that between cases where *one* agent makes a decision and performs an action (individual decision situations) and cases where *several* agent are involved (collective decision situations). It is not important whether these agents are persons or, say, corporations.

[11]Distinguishing between options and actions primarily serves to differentiate between the chosen option and the action subsequently performed. The two expressions allow us to talk about the same thing in two different ways: Actions are performed, options are instantiated (or realized)—by performing the corresponding actions. However, an agent faces *several* options but performs *one* action—or *at most one*, if inaction is not necessarily counted

normal conditions, an agent performs the action corresponding to the option they chose. Following an unwritten convention of normative ethics, I use "ϕ" (and sometimes ψ) for denoting options and actions. The context should make clear whether the one, the other, or both are meant.

Contexts and Consequences: Performing an action has *consequences*. An action's consequences are what, relative to some (often only implicitly assumed) context, would be the case if that action were performed but would not otherwise be the case.[12] Strictly speaking, *options* have consequences (only) in a derived sense, namely the consequences that their corresponding actions will or would have (again, relative to some context). That said, I will generally choose the less long-winded way of expression and not mention this every time.

The actual consequences of an action (usually) depend on a number of facts that obtain at the moment of action. We call those facts on which the actual consequences of the action depend, the *actual context* of a decision situation (at a point in time). We call the totality of facts (at a point in time) the *circumstances* (at that point in time). Figure 2.1 visualizes the relation between circumstances and context. However, the (actual) context might have been different. In that case, the consequences of some action might have been different *because* the (actual) context had been different. We call contexts that could be the case *contingent contexts*. Contingent contexts can play a crucial role in the normative assessment of actions, namely if what is considered *normatively relevant* is not the actual context but, for instance, the contexts that can or should be expected by the agent (more on this later). Figure 2.2 visualizes the relation between genuinely and actual contexts.

For the sake of simplicity (and because it corresponds to the practice in most parts of the debate), we will generally assume that contexts are static in relation to decision situations and during their resolution, i.e., they do not change during deliberation within a decision-making situation nor in the time between decision and action.[13]

While formal precision in describing decision situations is important for this project, their metaphysical and ontological nature is not. Accordingly,

as an 'act of omission'. Options are potential ways to act, while an action is the enactment of a chosen option, manifesting an event in the world. Sometimes things can go wrong, for example due to weakness of will, and an agent ends up performing an action that does not correspond to the option they actually chose beforehand.

[12] This characterization is somewhat simplified and will be examined in more detail later.

[13] However, this assumption cannot be maintained over the course of the project and will be relaxed when the time comes.

I focus on establishing a clear language and framework for analyzing and modeling them and refrain from engaging in metaphysical discussions. I would be willing to go so far as to assert that decision situations, in the context relevant to this thesis, can be conceptualized as some kind of structure involving (transient) states of affairs[14] and the above specific elements as building blocks.[15]

It's worth noting that decision situations may encompass additional elements, and in different contexts, they may be characterized differently. Further, one should be careful not to confuse the representation of a thing with the thing itself: The formal descriptions employed in this thesis are, in a sense, models *representing* decision situations, typically merely hypothetical ones. Often this differentiation doesn't matter, sometimes it does. Where I think that it matters, I try to be as explicit as possible.

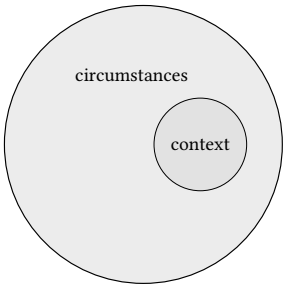


Figure 2.1: Circumstances and context. The actual context is a part of the actual circumstances.

With this conceptual groundwork laid, I suggest the following definition:

Definition 2.1 (Individual Decision Situation) *An* individual decision situation *is a situation in which a single agent is presented with multiple options to choose from, and within a given context each action that corresponds to the agent's options has an associated (actual) consequence.*

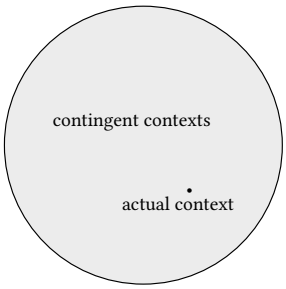


Figure 2.2: Set of potentially (normatively) relevant contexts. The actual context is one potentially relevant context, all other potentially relevant contexts are *merely* contingent.

Here is a description of an individual decision situation, a modified version of an example by Felix Pinkert (Pinkert 2015), which will later take a central role in this project:

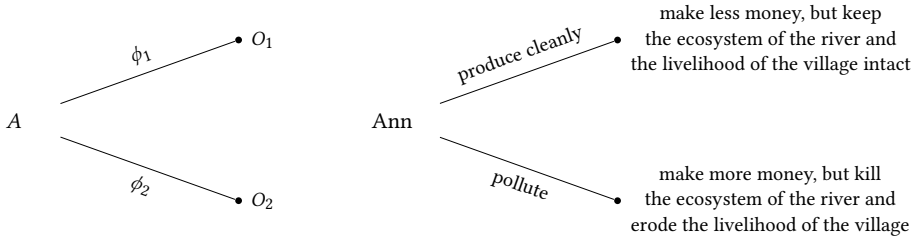
Case 2.1 (Factory) *Ann owns a factory near the river. Ann can either produce cleanly or pollute. Polluting would allow Ann to produce significantly cheaper and, as a result, make some extra cash which would enable her to afford another week of vacation at a diving resort. But polluting would also kill all the fish in the river and erode the livelihood of a village downstream.*

[14] A state of affairs is *transient* if, and only if, it obtains at one time and not another (cf. Textor 2021).

[15] Here, one may think of branching time models in the sense of A. N. Prior (1955; 1967), where options are possibilities to choose between potential futures that are ultimately overlapping sets of different kinds of propositions or states of affairs. Stit semantics seem to me to represent a particularly promising modeling of this idea (cf. Belnap, Perloff, and Xu 2001; Horty 2001; Horty and Belnap 1995).

Before I proceed, a word on the relationship between cases and decision situations: Cases are best understood as *descriptions* of decision situations. This means that we can easily jump between some case (description), which is a linguistic entity, and the decision situation described by it. Of course, this presupposes that the description is such that it actually describes a possible decision situation, i.e., that the description is coherent, consistent, and sufficiently complete. In particular, such a description must specify a decision situation in terms of the components outlined above and with regard to a relevant context. A proper case in that sense therefore warrants the *existence* of the described situation in a metaphysically 'lightweight', deflationary way: if there is a coherent, consistent, and sufficiently complete case description, then there is a corresponding decision situation (together with a relevant context) that is described by it. This goes well with the arguably very convincing view that not only actual but also conceptually possible[16] decision situations—let us call their descriptions "thought experiments"—are relevant to and a benchmark for normative theories. A view that certainly agrees with the practice of normative ethics (and rationality theory).[17]

Let us now furnish our toolbox with some representational devices. Sometimes it is handy to depict a decision situation—or rather its structure—visually. We can depict *minimal* individual situations of some agent A with some options, say, ϕ_1 and ϕ_2 with consequences O_1 and O_2, in terms of simple "decision trees". Here is a generic example on the left and an instance of Factory on the right:



Obviously, we can generalize this device to decision situations with an arbitrary number of options.

Having some formal notations at hand pays off later. Given a specific decision situation D of an agent A with options ϕ_1 to ϕ_n I call $\Phi = \{\phi_1, \ldots, \phi_n\}$ the *option space* of A.[18] I assume that a decision situation implies that a *choice*

[16] The reader is welcome to weaken the claim to nomologically possible situations. We could add to the list of conditions above—being coherent, consistent, and sufficiently complete—the condition of being in line with the laws of nature.

[17] Should this be doubted, I would refer to the myriad trolley examples and many other thought experiments that not only populate introductory courses but are also commonly elevated to the status of standards for correct moral assessments.

[18] I restrict my investigation to cases with finitely many options, contexts and (possible) consequences.

can be made, that is, in particular, that every agent always has at least two options, i.e., $|\Phi| \geq 2$. I also want to be able to represent dynamics and uncertainty later on. To this end, I allow more than one context per decision situation by defining a relevant set $\mathcal{C}$ of relevant contexts for a given situation D. What makes a context *relevant* is a question that demands an answer from a normative theory—I postpone it to the next subsection. Furthermore, we call the set of possible consequences (of the options of a decision situation D) $\mathcal{O} = \{O_1, \ldots, O_m\}$. As we assume that the outcomes are a function of action and context, we get $m = k \cdot n$ relevant outcome where k is the number of relevant contexts, i.e., $k := |\mathcal{C}|$, and n is the number of options. Accordingly, we can model this relationship between actions and contexts on the one side and the consequences on the other as an *outcome function* $\mathsf{Out} : \Phi \times \mathcal{C} \to \mathcal{O}$. Since we have no choice of contexts for singletons $\mathcal{C} = \{C\}$, we may, for the sake of simplicity, write $\mathsf{Out}_C : \Phi \to \mathcal{O}$ instead of $\mathsf{Out} : \Phi \times \{C\} \to \mathcal{O}$ in these cases. When more than one decision situation is under consideration, I sometimes use indices for disambiguation, e.g., I write $\mathcal{C}_D$ for the set of relevant contexts of a decision situation D or Φ_D for the option space of that decision situation; where there's no need, I leave them out.

It is practical and convenient to have a common agreement on formal entities representing a decision situation[19] as well (and not only of their constituents introduced above): Let D be an individual decision situation of agent A with options Φ, a corresponding outcome function Out, and a set of relevant contexts $\mathcal{C}$. D is represented by the tuple $\mathsf{D} := \langle A, \Phi, \mathcal{C}, \mathcal{O}, \mathsf{Out}\rangle$ with $\mathsf{Out} : \Phi \times \mathcal{C} \to \mathcal{O}$.[20] Occasionally, I will introduce a decision situation in a compressed way, for example like this: $\mathsf{D} := \langle A, \Phi, \mathsf{Out} : \Phi \times \mathcal{C} \to \mathcal{O}\rangle$. By explicating the signature of the outcome function within the tuple, I thereby explicitly introduce symbols for the relevant contexts $\mathcal{C}$ and the outcomes $\mathcal{O}$ into the discourse. Finally, let $\mathfrak{I}$ denote the set of all individual decision situations.

After these meta-comments and preliminary formalities, it is now time to start the reconstructive work.

2.2 Moral Theories and the Rightness Predicate

Now that there is a clear understanding of individual decision situations, we can turn to the normative questions to be asked about these situations. In particular, with respect to MH and Collectively Maximizing, we need a precise

[19] I will introduce *collective* decision situations in a similar way later in this book. For now, individual decision situations will do.

[20] Note that this is meant to establish an implicit naming convention: If we are considering two individual decision situations D_1 and D_2, we can easily switch to the level of their formal representation by switching to the corresponding tuples D_1 and D_2.

way of articulating assessments regarding the rightness of options or actions according to some moral theory. At base, moral theories are answers to a specific question, viz.,

(N) What is right for an agent to do (and why)?

Typically, moral theories answer this question in a principled way with regard to individual decision situations. Since we have a sufficiently clear conceptual framework of decision situations at hand, we can turn to the substantial, normative aspects of (N).

Before I turn to the family of moral theories that is central to this project, i.e., Maximizing Objective Act-Consequentialism (MOAC), I briefly introduce some normative terminology and a more formal way of writing precisely about moral assessments relative to some theory. A wide variety of interrelated normative concepts are relevant in the context of (N). For instance, how is the right thing to do related to what one *ought* to do? Or to what is wrong to do? And what does it mean for some outcome to be good or bad?

These and related questions have at least two dimensions. First, they can be about different kinds of *normative statuses* (or properties); second, they can be about the different things that can *have* these statuses and properties: agents, options, actions, intentions and motives, states of affairs, … In everyday language, the use of the corresponding terms is not precisely regulated. In the context of moral philosophical theorizing, the challenge is to find, on the one hand, theoretically fruitful characterizations and explications, but, on the other hand, characterizations and explications that are also sufficiently close to common usage. Otherwise, our moral philosophical theories might cease to be useful for our everyday discourses. It is thus a challenge typical for analytic philosophy.

The following characterizations are intended to conform to the ideal of analytic philosophy insofar as I make my understanding explicit. However, I am not claiming that my characterizations are fundamentally and metaphysically true—regardless of whether it makes sense to assert such claims at all, which I generally doubt. Instead, the picture drawn below is intended to be sufficiently close to that of the philosophical discourses relevant to this project while at the same time fruitful for my own purposes.

I propose to understand the concept of normative (or, in the context of this thesis, moral) status to be non-monolithic. We should distinguish between *deontic* and the *evaluative* statuses. In this book, I am concerned with the deontic status of options and actions and thus with whether, for example, a particular choice of an agent in a particular decision situation is right, wrong, neutral, forbidden, permitted, or obligatory. Furthermore, I limit myself to questions regarding rightness or wrongness, as this corresponds most closely to the literature on the CHALLENGE, and there does not appear to

be any unfair simplification of the matter resulting from this limitation. I do not exclude the possibility that there is a fundamental difference between rather *prescriptively* loaded notions, like obligatoriness, permissibility, and forbiddenness, on the one side, and rather *descriptive*-sounding[21] notions, like rightness and wrongness, on the other.[22]

The exact relation between these properties need not concern us at this moment, but here is a proposal based on a suggestion by Krister Bykvist (Bykvist 2003, pp. 45-46):

(1) An action ϕ is *obligatory* for an agent A in (a decision situation) D if and only if ϕ's outcome would be better than the outcome of any alternative action for A in D.

(2) An action ϕ is *right* for an agent A in (a decision situation) D if and only if ϕ's outcome would not be worse than the outcome of any alternative action for A in D.

(3) An action ϕ is *wrong* for an agent A in (a decision situation) D if and only if ϕ is not right for A in D.

Relevant here is not whether these statements are substantially correct. What is at issue is that these quotes emphasize that, for champions of MOAC, the scheme shown in Figure 2.3 typically applies: if there is only one right action, it is also obligatory (and if there are several right actions, it is obligatory to perform one of these); moreover, the sets of right (wrong) and of permitted (forbidden) acts are identical. I call this the Consequentialist Standard View. Thus, where necessary or useful, I feel free to switch between describing an action as right and describing it as permitted, on the one hand, and describing an action as wrong and describing it as forbidden, on the other, especially when interpreting specific quotations.

Like the Consequentialist Standard View, I, for the most part, ignore the possibility of morally neutral or genuinely conditional statuses. This is indeed

[21] Whether they are actually descriptive depends, obviously, on certain meta-ethical assumptions, first and foremost, moral realism. However, I do not commit to any such view here as it does not matter with respect to the CHALLENGE.

[22] I believe that understanding actions, or options, respectively, as the primary bearers of deontic statuses corresponds to our everyday linguistic practices. Especially in deontic logic, however, it is typical to use propositions or states of affairs as carriers instead. Accordingly, Op for a deontic operator $O\ldots$ then reads approximately "It ought to be the case that p". While I do not want to deny that there is a meaningful way of talking like that, there is evidence that some of the classical 'dilemmas' of deontic logic result from not referring to agency and actions instead, as stit semantics do (cf. most importantly Horty 2001). For this project, I settle on actions and options as primary, or at any rate essential, bearers of deontic statuses.

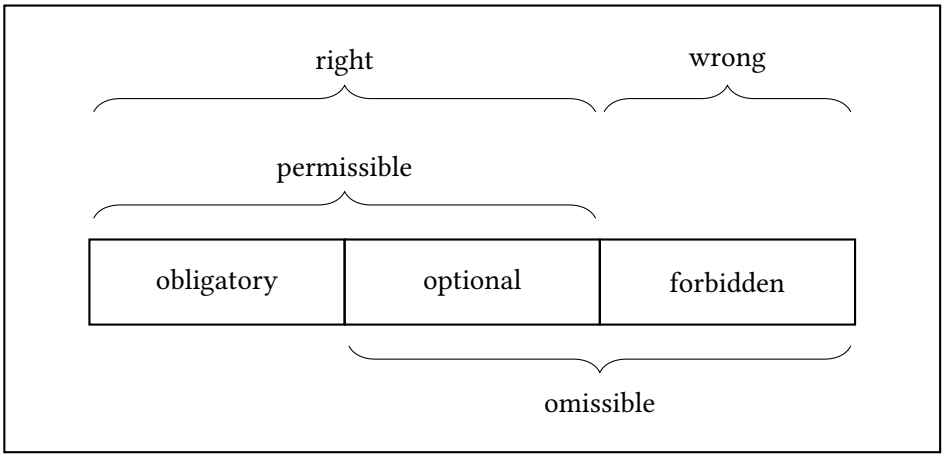


Figure 2.3: The relation between the different kinds of deontic statuses according to the Consequentialist Standard View (cf. McNamara and Van De Putte 2022, Fig. 3).

actually quite common. For instance, note that Bykvist's definitions above even exclude the possibility of such statuses. Neutral actions, i.e., actions that are neither right nor wrong, are excluded because all actions that are not right are, according to (3), wrong. For similar reasons, actions with *genuinely conditional* deontic status are excluded, i.e., actions that are right (or wrong) only if a certain condition holds, but neither wrong nor right independently from whether said condition holds or not. Assume there where an action that is genuinely conditionally moral status. Either such an action is right or it is not right—there is no third. If it is right, it has a non-conditional moral status. However, if the action is not right, then that action is wrong according to (3). Thus, in both cases, that action would have a non-conditional moral status and hence they would not have *genuinely* conditional deontic status. A contradiction.

However, we will see that consequentialists are well advised to make conceptual space for genuinely conditional deontic statuses, at least in the current form of their objective consequentialist framework. (The version of consequentialism that I propose later in this book does not need conceptual space for such exotic and esoteric statuses. But I don't want to get too far ahead of myself.)

Furthermore, there are so-called *evaluative* statuses, of which there are two kinds, namely *absolute* evaluative statuses—being good, being bad, and possibly being (evaluatively) neutral—and the *comparative* evaluative statuses—being better than, being worse than, and being equally good as. The latter statuses are always relational, i.e., relative to other options or actions. The following distinction proves useful for this project: Theories that ascribe

evaluative statuses are called *axiologies*; theories that ascribe (moral) deontic statuses are *moral theories*.[23] I will return to axiologies and their role within consequentialist theories later in this chapter.

We can now formally introduce moral theories as theories that—quite in the spirit of question (N) above—operate on individual decision situations and assign the options within their corresponding option space a deontic status in a principled way. This exclusive focus on individual decision situations may seem 'innocent', but against the backdrop of the collective character of the Challenge (i.e., of Troublemakers), this commonly accepted aspect of moral theories is significant for the present project. It is, therefore, worth being a little more explicit on this point. In this sense, I take it that the following principle[24] expresses a widely accepted but seldom explicated view:

Principle 2.1 (Methodological Individualism)
The primary bearers of deontic status are options of moral agents. Whatever has a deontic status and is not itself the option of a moral agent has this status merely in a derivative sense, that is, deontic status is a function of the deontic status of certain options of moral agents.

Methodological Individualism may raise various questions, the most pressing of which is arguably why options and not actions should be the primary bearers of deontic status. Although I believe that, in principle, both views—that either options or actions are the primary carriers—can be defended, I also believe that the question of the *appropriate* explication of the principle does not ultimately depend on some metaphysical truth, but on which perspective one adopts with which particular theoretic interest. In other words, the choice of the specific conceptual framework should ultimately also depend on its usefulness for the respective endeavor of theory building.

[23] Of course, there are also completely different conceptions of what a moral theory is and how moral theories relate to axiologies and possibly further kinds of theories. For example, Marc Timmons suggests that moral theories are composed of a "Theory of Right Conduct" and a "Theory of Value", which makes statements about both "Intrinsic Value" and "Moral Worth" (cf. Timmons 2001, pp. 9). For the present (consequentialist) project, however, a theory of intrinsic value (i.e., an axiology) is sufficient—and what Timmons calls a "Theory of Right Conduct" is, in my picture, a certain part of a moral theory that is typically characterized by a specific criterion of rightness (plus some relationships between further moral statuses, see above). But we will come to the details in a moment. What is important here is this: By no means do I want to offer a general analysis of the conception of moral theory. I do not want to commit myself to *any* one such conception. (Actually, I do not believe that it is worth searching for any *unique* such conception since different conceptions are differently suited for different purposes.)

[24] Originally, I thought I had come up with a wonderfully apt name for this principle. However, it turns out that there is an established principle by this name in the social sciences (cf. Heath 2020; Schumpeter 1908; Weber, Roth, and Wittich 1978) which actually has a close conceptual relationship to the principle proposed here, as it calls for social phenomena to be explained by appeal to individual actions.

If one aims to describe the observable world 'from the outside', then it is probably expedient to regard actions as the primary bearers of deontic status because options are, in a sense, not observable, are no things or events in the world. However, if one is interested in understanding decision situations 'from the inside' and wants to build and develop theories about what the right thing to do for an agent in a given decision situation is (cf. question (N) above), then it is options that arguably should be considered as such primary bearers. Actions (as the manifestations or realizations of options) then inherit their moral status directly from the options that correspond to them. If an action is right (or wrong), it is *because* the agent has realized a right (or wrong) option through this very action. Since my project indeed revolves around decision situations and focuses on what is right for an agent to do in that sense, it is essential to know which option is right and, hence, which action is right to perform. Thus, I suggest the version of Methodological Individualism given above to be adequate in this project's context.

Most importantly, in the context of the Challenge, Methodological Individualism tells us something relevant about combinations of actions: If a combination of actions (a combination that fails to qualify as an action in itself) has deontic status, then, arguably (in the absence of better candidates), this status is a function of the moral statuses of its parts, i.e., the individual actions that make up that combination. Note that one can embrace Methodological Individualism without committing to the view that combinations of actions can have (or even normally have) deontic status at all. All it suggests is that *if* some combinations have such status, then they have it in virtue of the deontic status of the individual actions that constitute these combinations. A simple principle in line with Methodological Individualism could be, for instance, that some combination of actions is right if all actions it contains are right; and that such a combination is wrong if some contained actions are wrong (we discuss such principles at the very end of this part in the context of Frank Jackson's approach to the Challenge, cf. Jackson 1987).

Further, Methodological Individualism does not rule out the existence of group agents nor that their actions have some kind of genuine deontic status. If they qualify as genuine agents (for elaborate accounts cf. List and Pettit 2011; Pettit and Schweikard 2006), they can populate individual decision situations. However, Methodological Individualism rules out that *mere* collectives, loose groups that do *not* qualify as agents, can be the ultimate addressees of a moral theory's guidance. If some combination of actions is made up of actions of individuals standing in such a loose connection and if that combination has some deontic status at all, then this deontic status is a function of the deontic statuses of the individual agents' actions. According to Methodological Individualism, thus, a moral theory is not allowed to assign deontic statuses to such combinations.

Since one could assume that not every moral theory has to be applicable to every decision situation, it can be useful to restrict one's reasoning about a moral theory T to a set $\mathfrak{I}_T$, the set of decision situations for which T has something to say, i.e., to T's *domain*.

It will come in handy later to have a shorthand for "given a decision situation D (of agent A with options Φ) and a set of relevant contexts $\mathcal{C}$ it is right (for A) to ϕ according to moral theory T". For this, we will use a rightness predicate R and write[25]

$$T, \mathsf{D}, \mathcal{C} \vDash \mathrm{R}\phi.$$

We can give a set-theoretic semantic for R. For this, we first define some useful sets. Let $\mathcal{D} \subseteq \mathfrak{I}$ be a set of individual decision situations. $\mathcal{C}_\mathcal{D}$ denotes the set of all the relevant contexts of the decision situations in $\mathcal{D}$ and $\Phi_\mathcal{D}$ denotes the set of all the options available in the decision situations in $\mathcal{D}$, i.e.,

$$\mathcal{C}_\mathcal{D} := \bigcup_{D \in \mathcal{D}} \mathcal{C}_D \quad , \quad \Phi_\mathcal{D} := \bigcup_{D \in \mathcal{D}} \Phi_D.$$

We can think of a moral theory T, then, in terms of a function $\mathcal{T}$ that maps a decision situation D from T's domain $\mathfrak{I}_T$ with an option space Φ_D and fitting context $C \in \mathcal{C}_D$ into a set of right actions $\Phi_T \subseteq \Phi_D$, i.e.,[26]

$$\mathcal{T} \,:\, \mathfrak{I}_T \times \mathcal{C}_{\mathfrak{I}_T} \to \Phi_{\mathfrak{I}_T}$$

with

$$\mathcal{T}(D, \mathcal{C}) := \mathcal{T}(\mathsf{D}, \mathcal{C}) = \mathcal{T}\left(\langle A, \Phi_D, \mathsf{Out} :\ \Phi_D \times \mathcal{C}_D \to \mathcal{O} \rangle, \mathcal{C}\right) \subseteq \Phi_D$$

where $\mathcal{C} \subseteq \mathcal{C}_D$ and

$$\mathcal{T}(D, \mathcal{C}) := \{\, \phi \in \Phi_D \mid \phi \text{ is right for } A \text{ given } \mathcal{C} \text{ according to } T \,\}.$$

Note that the function $\mathcal{T}$ is undefined for decision situations outside of T's domain or irrelevant contexts, i.e., for a context $\mathcal{C} \nsubseteq \mathcal{C}_D$.

We can now give the shorter, but equivalent, semantical definition

$$T, D, \mathcal{C} \vDash \mathrm{R}\phi \quad \text{ if and only if } \quad \phi \in \mathcal{T}(D, \mathcal{C}).$$

For the time being, we can leave open the question of whether there is more to be said about a wrongness predicate W beyond indicating that an action is not right—that is, whether we should accept

$$T, D, \mathcal{C} \vDash \mathrm{W}\phi \quad \text{ if and only if } \quad T, D, \mathcal{C} \nvDash \mathrm{R}\phi$$

[25] If only one context is relevant (which is the case for the most part of my project), I will simply mention the individual context instead of the singleton.

[26] Again, if only one context is relevant, I will simply mention the individual context instead of the singleton.

and thus whether we can simply define

$$T, D, C \models \mathrm{W}\phi \quad \text{if and only if} \quad \phi \notin \mathcal{T}(D, C).$$

If the Consequentialist Standard View is true, there is nothing to be said against this simple formalism (though later, in the context of collective decision situations, I will raise doubts about this assumption).

At this point, it might seem that entailing a predicate of rightness is all we should and could expect from moral theory. While this is true in a certain sense, such a predicate of rightness entails much more than merely a *criterion* of rightness. What may sound a bit cryptic at first will be explained in the next section in much more detail. For the time being, however, it is sufficient to remember the criterion of rightness, which has already been roughly outlined in the introduction and which, so to speak, plays a main role in this work. Recall the Maximizing Objective Criterion of Rightness:

Criterion 1.1 (MOCOR – tentative) *An action is right if and only if there is no alternative action that would lead to better consequences.*

We can now capture this a bit more precisely as

Principle 2.2 (MOCOR) *Let D be an individual decision situation involving an agent A with a set of options Φ and with actual context C. An action $\phi \in \Phi$ is right for A in D given C if and only if, relative to C, there is no alternative action $\phi' \in \Phi$ with better consequences than ϕ.*

As already mentioned in the introduction, any moral theory embracing this criterion of rightness is, by definition, part of the family of theories I call Maximizing Objective Act-Consequentialism (MOAC). But what is the specific function that makes up $\mathcal{T}_{\text{MOCOR}}(D, C)$? I cannot provide a formal condition for it at this stage, as the framework developed thus far lacks a crucial component: a method for morally evaluating the consequences of actions and comparing them. In the following section, which concludes this chapter, this gap will be addressed.

2.3 MOAC and Its Three Modules

Looking back, we find at least three aspects a completely specified criterion of rightness could operate on: the agent, their options, and their consequences. These three aspects correspond to the traditional triad of 'pure' families of moral theories: the family of virtue ethics focuses on the agent, especially their character; for the family of deontological theories, the moral status of options and actions is, first and foremost, a function of the actions themselves,

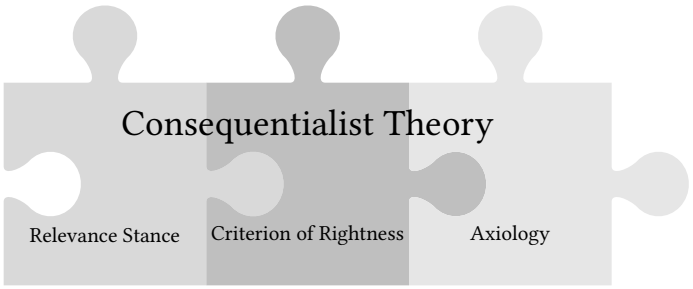


Figure 2.4: The modular nature of consequentialist theories. They consist of a view regarding which contexts are relevant, a criterion of rightness, and an axiology. As discussed in the following, MOAC consists of Objectivism as Relevance Stance, Maximization as Criterion of Rightness, and an axiology that allows for ranking outcomes (as discussed in the next section).

their types and properties, i.e., their universalizability and adherence to certain (universalizable) rules; for the final family, consequentialism, the moral status is determined *solely* by the consequences of the options. Naturally, there can be hybrid theories. While admittedly painted with a broad brush, this distinction suffices for demarcation purposes.

Since consequentialist theories know no limitations as to their applicability, their domain is the set of all decision situations.[27] It is thus appropriate to characterize consequentialist theories as exactly those moral theories that embrace a particular fundamental meta-normative principle, viz.:

Principle 2.3 (Essence of Consequentialism) *For every individual decision situation D and associated set of relevant contexts C: the moral status of some options available in D is solely determined by the morally relevant qualities of the consequences of that option (typically relative to the morally relevant qualities of the consequences of its alternatives) as determined by C.*

This principle raises three key questions requiring clarification: Which contexts qualify as relevant? What determines the rightness of an option relative to these contexts? And what defines the moral quality of consequences? I propose, therefore, that a comprehensive consequentialist theory comprises three modules, each addressing one of these questions (cf. Figure 2.4): First, the *relevance stance* specifies which contexts are relevant. Since context, together with the action performed, determines the consequences of an action, this view directly defines which consequences are to be considered. Second,

[27] This statement should be taken with a grain of salt. I am convinced that it represents the default view of the vast majority of consequentialists, even if this is seldom made explicit (at least, I am unaware of any clear statement to that effect). Furthermore, doubts could be raised about such a perspective, as I will argue in the final chapter of this section. This question is relevant to my project, as I will later present consequentialism's self-image as a 'deonticely complete' theory as a motivational foundation.

a *criterion of rightness* establishes the deontic status of options based on the moral quality of their consequences, typically by evaluating the moral qualities of an option's consequences in relation to those of alternative options. Third and finally, a background theory of (non-instrumental) values, i.e., an *axiology*, determines the moral qualities (or value) of these consequences. In essence, axiologies place moral value on the workbench of consequentialists, the criterion of rightness dictates how these values influence the normative status of options and actions based on their consequences, and the relevance stance specifies which consequences are pertinent for this assessment.

The assertion that consequentialist theories are fundamentally defined by three modules, with the criterion of rightness being just one of them, seems to conflict with earlier statements in this book. Up until now, I maintained that a moral theory falls under MOAC if and only if it adheres solely to *one specific* criterion of rightness, namely MOCOR. Nevertheless, this is not an inconsistency but rather an ambiguity as MOCOR is, in fact, a composite of both a criterion of rightness (in the "module sense" as previously introduced) that I shall call "Maximization"—which corresponds to the "M" in "MOCOR"—and a relevance stance that I shall call the "Objective View"—referred to by the "O" in "MOCOR". The third module, however, the axiological background theory, is deliberately left unspecified because this is where the various MOAC theories diverge. Thus, MOAC is just one family of partially but not fully defined consequentialist theories.

A detailed step-by-step introduction to the three modules is invaluable for my project. This is particularly true concerning the axiological module, because one specific part of the axiological background theories that concerns questions regarding the aggregation of values will play a central role in the later stages of my project (while the other part, concerning what has inherent value in the first place, does not matter at all for my purposes).

2.3.1 Relevance Stances

What I call a "relevance stance" is a view regarding the question of which perspective on decision situations is relevant for their normative assessment. There are two kinds of relevant stances that are prominent in the consequentialist camp.

One is *objective*, in the sense that it looks at the situation from the outside. Intuitively but misleadingly formulated, this is supposed to express that the moral status of an action depends solely on what is actually the case. This statement is misleading because one can, of course, claim that the agent's having such-and-such beliefs about the situation (and about their options/actions and their probable consequences etc.) is also a fact—something that is actually the case. However, this is precisely what is *not* considered a relevant fact

according to that view. Instead, epistemic and doxastic aspects are deemed irrelevant.[28]

To express the idea more precisely, we can thus resort to the concept of the *actual context* of a decision situation. We have introduced it as that specific part of the situation's circumstances on which, given a particular action, its consequences depend (cf. page 16). However, because the consequences of an action generally[29] do not depend on the agent's beliefs, desires, or other aspects of their mental life, they typically[30] fall outside this context. We therefore define:

View 2.1 (Objective View) *The moral status of an action within a decision situation depends solely on the actual context of the decision situation.*

Under the Objective View, thus, the agent is only the 'executive organ' of the decision-making situation. It does not depend on their convictions, principles, attitudes, dispositions, nor on their particular (epistemic) history that led them to their situation at this point in time. It is entirely irrelevant what they have evidence for, what they are justified to believe, how strongly they are justified to believe, etc. Thus, according to the objective stance, each decision situation has only one *relevant* context, namely the *actual* context. This is the objective relevance stance.

The second kind of stance is *subjective* in the sense that as these stances assign relevance to epistemic aspects. Accordingly, they are agent-relative. These stances, thus, make the moral status of an agent's actions also a function of, for instance, the agent's beliefs or what the agent ought to believe and so on. According to subjective stances, there is usually a multitude of relevant contexts, for instance, all the contexts that the agent actually considers possible or that the agent *may* consider possible according to some normative *epistemic* standard (here, it depends very much on the details of the respective subjective stance). Based on a distinction made by Michael J. Zimmerman (cf. 2014), I hence suggest to differentiate between at least the following two subjective stances:

[28]For this reason, I personally would prefer to frame the following distinction as one between non-epistemic and epistemic stances. However, at least calling the non-epistemic stance "objective" is rather the default in the consequentialist debate (cf. Railton 1984; Sinnott-Armstrong 2022), and I will adhere to this convention to avoid unnecessary divergence.

[29]Exceptions are conceivable. Imagine that Timo, who has a severe headache, can either take a mysterious drug, the "truly potentially effective placebo", or not. If he takes the drug, it will bring relief to Timo if and only if he believes it to do so. In this case, the consequences of Timo's drug-taking depend objectively on Timo's beliefs. Similar dependencies between outcomes and an agent's mental states appear in cases like Kavka's toxin puzzle (Kavka 1983). In such rather exotic cases, the agent's mental states (or the mental states of other persons) belong to the actual context.

[30]Cf. Footnote 29 For readability, I will omit similar qualifications from here on.

View 2.2 (Subjective View) *The moral status of an agent's action within a decision situation depends solely on what contexts the agent believes to be possible in the situation.*

View 2.3 (Prospective View) *The moral status of an action within a decision situation depends solely on the beliefs that a epistemically rational agent would have about the possible contexts given the epistemic history of the agent.*

We need not dive into the details of these views in the context of my project since we are primarily concerned with defending MOAC theories and, thus, by definition, objective consequentialist theories. Nevertheless, the Prospective View plays a role in the context of the thesis, on the one hand, in reconstructing the CHALLENGE, and on the other hand, because I want to argue in the second part of my thesis that MOAC adherents can learn something important from it.[31]

One can take two positions with respect to the validity of these general perspectives: either one can see it as a dispute concerning the question of which view is 'really correct'. Then subjective/epistemic and objective/non-epistemic notions are *mutually* exclusive. Alternatively, one can see them as complementary, as two justified perspectives that together offer a richer moral framework. The complementary view can take several forms. For instance, one can disambiguate the same moral predicate (for instance, being right) subjectively and objectively or advocate an objective view for some predicates (say, being right) and a subjective view for others (say, being obligatory). According to the latter complementary view, one can hold that the question of what an agent *ought* to do aims at an answer that should be action-guiding and therefore formulated subjectively, while at the same time one can hold an objective view for moral *rightness* (cf. Andrić 2013; Ord 2005). In contrast, according to the first complementary view, one could hold, for example, that sometimes we are concerned with subjective rightness, while in other contexts we are interested in objective rightness. Then objective rightness arguably equals the 'epistemic limit case' of subjective rightness. Derek Parfit once expressed such a view (Parfit 1988, p. 2): "if, when acting, we know all the relevant facts", then the "two kinds of rightness [...] coincide". Somewhat more precisely,[32] we can capture the idea like this:

[31] The Subjective View is not generally considered particularly plausible anyway, see, for example, Jackson 1991. It has been included here only for contrastive purposes.

[32] However, the whole idea sketched here presupposes that knowing something implies absolute certainty and thus a credence of 1. Such an assumption might actually be implausibly strong and, thus, might be doubted for good reasons (cf. MacFarlane 2023; Williamson 2002). Under such a 'softer' concept of knowledge, an agent could then very well know that an option ϕ has the best consequences, and yet the expected value of $\neg\phi$ could still be greater than the value of ϕ. However, such concepts of knowledge can be set aside here: Where I bring the following principle (Epistemic Limes) into play later, I need it in order to be able to

Principle 2.4 (Epistemic Limes) *Let T_o be an adequate objective (i.e., non-epistemic) moral theory and let T_s be an adequate subjective (i.e., epistemic) moral theory. For a given decision situation involving an agent A with a set of options Φ and a set of relevant contexts C: If A knows all relevant facts and has no incorrect relevant beliefs, then*

$$T_o, D, C \vDash \mathrm{R}\phi \quad \textit{if and only if} \quad T_s, D, C \vDash \mathrm{R}\phi.$$

In other words, according to Epistemic Limes, adequate objective and adequate subjective moral theories are extensionally equivalent under the assumption of perfect epistemic situatedness of the involved agent.

In the context of this work, we do not really need to commit ourselves in this regard. But where subjective and objective consequentialist theories play a role in the course of the following chapters, it makes sense to take a complementary perspective, according to which the objective view is an ideal case of the subjective one, i.e., a case in which it is presupposed that the agent knows everything that is relevant and believes no falsehoods.

2.3.2 Criteria of Rightness

The second module of a consequentialist theory is a criterion of rightness. This criterion defines which deontic statuses are assigned to options within the decision situations in the domain of the theory (and why they are assigned). In principle, this could serve as a criterion for determining what is right or wrong (or what is forbidden, permissible, etc.), with all other properties then inferred implicitly based on the specific relationships among the various deontic statuses (for example, as outlined in the Consequentialist Standard View above). Without loss of generality, my project limits itself to criteria of rightness, if only because this corresponds with the consequentialist tradition, especially with an eye on the literature regarding the CHALLENGE.

A criterion of rightness is both classificatory and explanatory. It is *classificatory* in the sense that it classifies certain options as right; it is *explanatory* insofar as it explains *why*[33] an option is right, i.e., it is not just the statement of contingent extensional equivalence:

argue about certain parts of the literature in which knowledge is presupposed in the context of the discussion of the CHALLENGE. Either these contributions are to be read in such a way that knowledge *implies* certainty, or they are not. I write assuming the first case, and only then do I need Epistemic Limes. If a 'softer' concept of knowledge is meant, one in which knowledge does *not* imply credence of 1, my general considerations on uncertainty and underdetermination can be applied.

[33] Accordingly, the typical "if and only if" formulation in such explanatory definitions should *not* be interpreted as simple material biconditional. Such a reading would entail that a statement like "ϕ is right if and only if ϕ has property F" was semantically equivalent to the statement "ϕ has property F if and only if ϕ is right". However, this is contrary to the intended meaning. Explanatory definitions like the first statement are meant to express that ϕ is right

Definition 2.2 (Criterion of Rightness) *A criterion of rightness is a classifying and explanatory statement that for each individual decision situation D in a domain $\mathfrak{I}_T \subseteq \mathfrak{I}$ specifies a sufficient and necessary condition under which the options of the agent in D are right, and why.*

Here is MOAC's criterion of rightness in this technical, isolated sense:

Principle 2.5 (Maximization) *For a given decision situation involving an agent A with a set of options Φ and a set of relevant contexts $\mathcal{C}$: An action $\phi \in \Phi$ is right for A if and only if, relative to $\mathcal{C}$, there is no alternative action $\phi' \in \Phi$ with better consequences than ϕ.*

Note that this formulation intentionally leaves open which contexts are relevant. As such, it stands for one clearly separated module of MOAC theories. If we combine the objective stance, the first characteristic module of MOAC, with Maximization, we get the already known

Principle 2.2 (MOCOR) *Let D be an individual decision situation involving an agent A with a set of options Φ and with actual context C. An action $\phi \in \Phi$ is right for A in D given C if and only if, relative to C, there is no alternative action $\phi' \in \Phi$ with better consequences than ϕ.*

In this sense, it becomes clear that MOCOR is indeed a composite of the objective stance and Maximization as a criterion of rightness. It is, therefore, absolutely correct to say that MOAC theories are precisely those consequentialist theories that embrace MOCOR, and, at the same time, to say that this determines two out of three consequentialist modules. With this, we can finally turn to the third module.

2.3.3 Axiological Background Theories

Finally,there is the axiological module of consequentialist theories. Axiology is the philosophical study of value. Axiological theories are theories about what things have value and how much value they have, whereby "value" refers to value in an intrinsic, non-instrumental sense, sometimes also called "final value" (cf. Schroeder 2021). Traditionally, consequentialist theories adopt so-called welfarist axiologies, i.e., accounts holding that nothing but welfare (or well-being) matters morally.

In the tradition established by Parfit (Heathwood 2020; Parfit 1984), we can distinguish three types of welfarist axiologies: Hedonism, Preference

because ϕ has property F, i.e., that F is, in some sense, more fundamental than being right and, thus, explains *why* ϕ is right, or is what *makes* ϕ right—and not the other way around. To improve the precision, thus, one could use "if and only if, and if, then because" (or the more common "if and only if, and because", cf. Timmons 2001) to explicate the explanatory direction. However, for the sake of brevity, the shorter version will be used in this document, always implying the more elaborate interpretation for explanatory definitions.

Theories, and Objective List Accounts. I do not intend to engage with this debate here since the CHALLENGE is independent of any account of welfare, even though some of the examples I explore might presuppose a specific axiology. These could, however, be rewritten in a way that works with any other plausible axiology—but I do not intend to enter the question of what makes an axiology plausible, either. The important thing is that one does not have to commit to any specific axiology at all when discussing the CHALLENGE. Whichever axiology might be correct (whatever that means), the CHALLENGE arises independently from that question. Thus, I follow a path other consequentialists have taken earlier: I assume that the question of what account of inherent value to choose is settled, or at least that it can be left to others. In line with tradition, however, I presume a welfarist account in this book. Let us call these types of theories axiologies in the narrow sense.

There is another, *broader* sense of axiologies, according to which they, in addition to the narrow sense, are also concerned with method(s) of aggregation of value. After all, consequentialists are, in a sense, more interested in the *value of consequences* than in how well some specific individual is off, where, typically[34], the value of consequences is a function of how well individuals are off those consequences. Consequentialists hence typically need an account of what makes some consequences better than others. In other words, consequentialist frameworks need methods that lift value from the level of individuals to the level of consequences or even the level of *sets* of such consequences (if we take, for instance, a subjective stance and thus allow for several possible consequences). Throughout this book, whenever axiologies are mentioned, it should be apparent from the context whether I refer to them in the narrow or the broad sense.

It is convenient to have names for the two parts of the axiologies in a broader sense:

Grounding Part: The central question to be answered by the grounding part is: What has how much value (and why)? This part identifies *the value within* consequences, i.e., the welfare of individuals. I call this part the grounding part since it grounds the value of (sets of) consequences: a (set of) consequences has a certain value *because* of the value within those consequences. This part corresponds to axiologies in the narrow sense.

Aggregation Part: The central question to be answered by the aggregation part is: How good is a particular consequence or set of possible consequences? The aggregation part is thus instrumental in illustrating the translation of the value within consequences—as defined by the

[34]There are attempts to develop non-aggregative consequentialist theories (cf. Gustafsson 2021). These approaches are irrelevant in the context of this thesis.

> grounding part—into the overall value of the consequences (or sets of consequences). Such lifting of value from the individual to (sets of) consequences necessitates at least two forms of aggregation: interpersonal and intertemporal. Here, *interpersonal* aggregation refers to the methodologies for amalgamating values across individuals, while *intertemporal* aggregation denotes the methodologies for totaling up values over time. In general, simple summation is the favored approach for these forms of aggregation, particularly when the scope is confined to finite populations and temporal spans.[35] Furthermore, when it comes to 'interpossible' aggregation, which is concerned with 'uplifting' the value of consequences to the value of sets of consequences (typically representing possible consequences), subjective consequentialists often compute the sum of consequence values, each weighted by the subjective probability of its occurrence, effectively calculating *expected values*. This approach finds its roots in decision theory (cf. Jackson 1991; von Neumann and Morgenstern 1947).

Neither the grounding nor the aggregation part plays a significant role when it comes to *formulating* the CHALLENGE. However, as I am going to argue in the second part of this thesis, the aggregational part does play a crucial part in *solving* the CHALLENGE. Since this part of the book tries primarily to better understand the CHALLENGE, for now, we only need to be in a position to decide what MOAC recommends in certain situations and to evaluate whether this satisfies MOAC's own expectations. We can, therefore, simply postulate that outcomes of certain actions and combinations of actions have specific values, without the need to specify any concrete axiology. What constitutes the moral quality of outcomes,and how it does so, does not matter for the 'valuative profile' of the cases.

Thus, all that matters is that cases with certain 'valuative profiles' exist, but this can be plausibilized pretheoretically, purely based on common sense. Recall Factory from above. That it is terribly bad to destroy the river's ecosystem and, in doing so, the livelihood of the village downstream should be beyond doubt and arguably is *ceteris paribus* true with any plausible axiology; the same holds for the assumption that the loss of jobs is awful in a country with underdeveloped social safety infrastructure. It should also be clear that, regardless of our precise notion of what makes one state of affairs better than another, it is better if none of the harmful consequences occur than if they all do. Further, we will certainly agree that a situation where only some of the bad consequences occur is better than the bad extreme and worse than the good extreme. This is all we need to formulate the CHALLENGE—and axiologies that fail to account for these judgments would, by all

[35] For infinite time horizons, one typically introduces a discount factor such that moments further in the future count less than temporally closer moments.

appearances, have to be rejected, anyway. This is why the grounding part actually remains irrelevant for this endeavor. As a result, it is justified to remain at the level of MOAC —that is, to work with a *family* of moral theories rather than narrowing down to a specific theory, such as classical Utilitarianism. Although members of this family may differ (at least) in the grounding component of their axiological framework, this variation simply does not affect their susceptibility to the CHALLENGE.

Before we put all three modules together and pinpoint the essence of consequentialist theories, it is worth expanding the formal framework developed so far to include axiologies and assessments. This will pay off in the second, constructive part of the thesis.

Even though MOCOR, strictly speaking, only requires an axiology that allows us to distinguish the consequences with the best qualities from all the other consequences, it is typically assumed that the axiology implies a *valuation function*, i.e., a function $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$ assigning arbitrary outcomes from $\mathcal{W} := \bigcup_{D \in \mathrm{I}} \mathcal{O}_D$ a value from some value space $\mathcal{V}$. Such a value space is derived from the grounding component of an arbitrary axiology and is typically a mathematical space—that is, a set or domain of values. along with certain operations, such as addition and scalar multiplication. For some decision situation D, let us call $\mathsf{Val}(\mathcal{O}_D) := \{\, \mathsf{Val}(O) \in \mathcal{V} \mid O \in \mathcal{O}_D \,\}$ the *valuative profile* of D.[36]

Consequentialists typically expect that all the outcomes of a decision situation can be ranked, i.e., that there exists a total order[37] over the valuative profile of each individual decision situation. In practice, the typical choice[38] of value spaces implies a total order over the *entire* space, which induces, of

[36] For a set $X' \subseteq X$ and a function $f : X \to Y$, $f(X') \subseteq Y$ denotes the *image* of X under f, i.e., the set $\{\, f(x) \in Y \mid x \in X' \,\}$. This allows us to introduce some properties of functions that I assume to be known within the context of my project, namely, that a function is called *injective* if and only if for every $x, x' \in X$, $f(x) = f(x')$ implies $x = x'$; that a function is called *surjective* if and only if $f(X) = Y$; and that a function is called *bijective* (a so-called bijection) if it is a 1-1 mapping, i.e., it is both injective and surjective.

[37] A *total order* $\leq$ on a set X is a binary relation that satisfies the following properties for all $x, y, z \in X$:

Reflexivity: $x \leq x$

Transitivity: if $x \leq y$ and $y \leq z$, then $x \leq z$

Antisymmetry: if $x \leq y$ and $x \leq y$, then $x = y$

Totality: $x \leq y$ or $y \leq x$

[38] Typically, most consequentialist (often implicitly) assume that $\mathcal{V} = \mathbb{R}$ and that the order is just the standard less-equal relation over the reals, with Fehige 1995 being a notable exception, using a real coordinate space, thus setting $\mathcal{V} = \mathbb{R}^n$, and suggesting a lexicographic order. (Strictly speaking, this is a somewhat oversimplified way of putting what consequentialists typically do here; it might be more accurate to say that they assume that $\mathcal{V}$ is *order-isomorphic* to $\mathbb{R}$—or $\mathbb{R}^n$ in Fehige's case.)

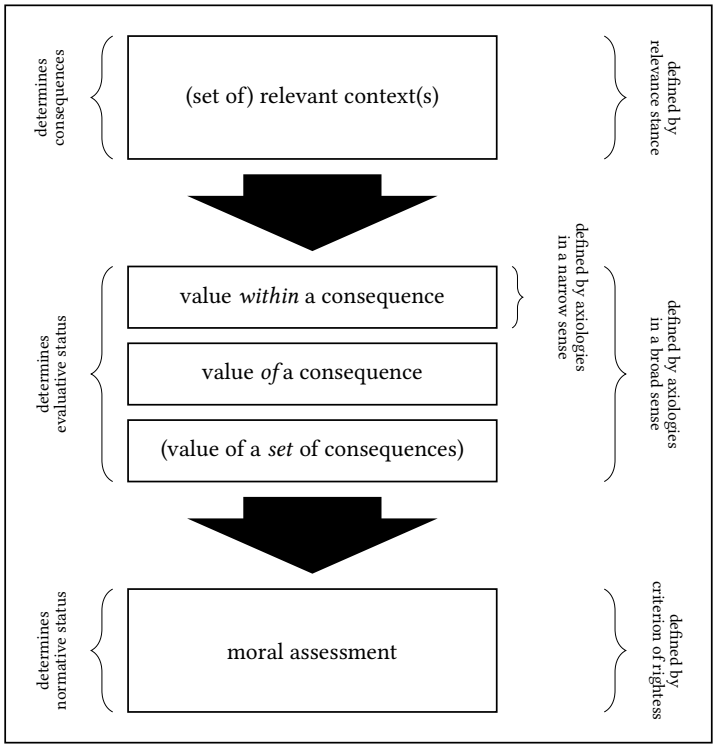


Figure 2.5: How the modules of a consequentialist theory interlink to arrive at moral assessments.

course, a total order over the valuative profile of each individual decision situations by definition. Even though, in principle, weaker constraints could be considered—and actually will be considered in the context of this project[39]—this assumption of the existence of total orders for arbitrary decision situations reflects quite common, often only implicitly made consequentialist assumptions. Accordingly, for the remainder of this part of the thesis, I presuppose the existence of a total order $\leq_{\mathcal{O}_D}$ over $\mathsf{Val}(\mathcal{O}_D)$ for arbitrary $D \in \mathfrak{I}$.

Finally, for an option ϕ in the option space Φ, an outcome function Out, and a relevant context C of some given individual decision situation, let $\mathsf{Val}_C(\phi)$ be agreed upon as an abbreviated notation for $\mathsf{Val}(\mathsf{Out}_C(\phi))$.

2.3.4 Putting It All Together

Finally, we can put all the modules together and arrive at the *consequentialist assessment pipeline*, consisting of three steps: first, the selection of the rel-

[39] MOCOR also works with partial incommensurability as long as there always is a 'best outcome', i.e., with partial orders plus some *completeness* property that, for a given $\mathcal{O}_D \subsetneq \mathcal{W}$, guarantees the existence of a greatest element in $V(\mathcal{O}_D)$. We will later encounter such a conception formulated in terms of not being dominated, cf. Horty 2001. For now, we can just ignore that possibility and presume local totality, i.e., the existence of a total order on $V(\mathcal{O}_D)$ for every D.

evant context by the relevance stance, yielding, together with the decision situation under consideration, the consequences; second, the evaluation of these consequence, in accordance with the axiology; and finally, the moral assessment as a function of the evaluative status of the consequences relative to the relevant context as defined by the criterion of rightness. Figure 2.5 illustrates this pipeline. We can now formulate a formal criterion for qualifying as an objective consequentialist theory:

Definition 2.3 (Objective Consequentialist Theory (formal)) *T is an* objective consequentialist theory *if and only if it embraces an axiological sub-theory T_{Ax} with a valuation function $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$ and an objective consequentialist criterion of rightness T_{CoR} such that, for all decision situations $D \in \mathfrak{I}$ and for all $\phi \in \Phi_D : D, C \vDash_T \mathrm{R}\phi$ if and only if $T_{\mathrm{CoR}}(\phi)$.*

A criterion of rightness T_{CoR} is objective consequentialist *if and only if, for all $D \in \mathfrak{I}$ with $\mathsf{D} := \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O} \rangle$ (with D's actual context C) T_{CoR} corresponds to a predicate $\chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}$ such that for all $\phi \in \Phi$:*

$$D, C \vDash_T \mathrm{R}\phi \quad \textit{if and only if} \quad \chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}(\mathsf{Val}(\mathsf{Out}_C(\phi))).$$

That is to say, for a moral theory T to be objective consequentialist, the rightness predicate implied by T's criterion of rightness is extensionally equivalent to a predicate ranging only over the values of options in a decision situation D and that, thus, solely operates on the valuative profile of D relative to the actual context C, i.e., on $\mathsf{Val}(\mathcal{O}_{D,C})$.[40] In other words: What makes an action right according to an objective consequentialist theory T *can only* be the moral quality of the consequences of this action (according to the axiological sub-theory of T, i.e., T_{Ax}) in comparison to the moral qualities of the consequences of the other options—and nothing else. This gives us a sharp criterion distinguishing objective consequentialist theories from other theories.

[40] In set theory, *characteristic functions* are functions that indicate the membership of elements in a particular set. The characteristic function of a subset A of a universal set U—in set theory, a universal set is a set that contains all objects considered in a particular context—is defined as follows:

$$\chi_A(x) = \begin{cases} \top & \text{if } x \in A, \\ \bot & \text{otherwise (i.e., if } x \in U \setminus A). \end{cases}$$

Mathematically, the characteristic function of a set A, often denoted by "χ_A", is defined from the universal set U to the Boolean domain (or Boolean set) $\mathbb{B} = \{\bot, \top\}$ (or to the binary set $\{0, 1\}$), so $\chi_a : U \to \mathbb{B}$. So one can freely switch back and forth between the predicate "… is an element of A" and the characteristic function associated with A.

Strictly speaking, thus, the predicates or functions characterized here are the characteristic functions of the set of permissible options relative to the moral theory under consideration, the decision situations, and their actual contexts.

Obviously, MOAC theories are objective consequentialist theories, not only by name but also according to this definition. In other words, we can state the relevant predicate of MOCOR explicitly as:

$$\chi^{\mathsf{Val}}_{\text{MOCOR},D,C}(\mathsf{Val}(\mathsf{Out}_{D,C}(\phi))) \quad \text{if and only if} \quad \forall\, \psi \in \Phi_D\colon \mathsf{Val}(\mathsf{Out}_{D,C}(\psi)) \leq \mathsf{Val}(\mathsf{Out}_{D,C}(\phi))$$

which is logically equivalent to the slightly shorter[41]

$$\chi^{\mathsf{Val}}_{\text{MOCOR},D,C}(\mathsf{Val}(\mathsf{Out}_{D,C}(\phi))) \quad \text{if and only if} \quad \phi \in \operatorname*{arg\,max}_{\phi' \in \Phi_D} \mathsf{Val}(\mathsf{Out}_{D,C}(\phi'))$$

That we can specify such a function for some moral theory T proves that, in the relevant sense, T is an objective consequentialist theory. We will see, however, that we can by no means specify such a function for all modifications of MOCOR that have been proposed in reaction to the CHALLENGE. Since this project subscribes to the search for an objective consequentialist solution to the CHALLENGE, Definition 2.3 will function as an important building block of a central criterion of adequacy for this thesis.

Before we turn to the reconstructive task of anchoring the CHALLENGE in its strongest form in the literature—and later the formulation of criteria for assessing moral theories as proposed solutions to the CHALLENGE—we briefly return to the question of possible *wrongness* predicates, linking back to the discussion above regarding the Consequentialist Standard View (cf. Section 2.2, page 21). Although we postpone the discussion of the pros and cons of the following two possible explications of MOAC's wrongness predicate until later, when they become relevant within the context of the second part of this project, we take the opportunity, while we are thinking about the rightness predicate, to think also about the concrete form of the wrongness predicate (or better, predicates).

The first variant, W_s, is the *easy way*. We simply define wrongness as a shorthand for not being right, i.e.,

$$D,C \models_T W_s\phi \quad \text{if and only if} \quad D,C \not\models_T \mathrm{R}\phi$$

We can translate this back to

$$D,C \models_T W_s\phi \quad \text{if and only if} \quad \neg\chi^{\mathsf{Val}}_{T,D,C}(\mathsf{Out}_C(\phi)).$$

[41] The operator arg max, short for "arguments of the maxima", is an operator used to identify the input value(s) at which a given function reaches its maximum output. Formally, if $f : X \to Y$ is a function that maps elements of a set X to elements of some totally ordered set Y, then, for some $X' \subseteq X$, $\operatorname{arg\,max}_{x \in X'} f(x)$ denotes the subset of X' containing those elements x for which $f(x)$ is maximized. In other words, $\operatorname{arg\,max}_{x \in X'} f(x) := \{\, x \in X' \mid f(x) \geq f(x') \text{ for all } x' \in X' \,\}$. The arg max operation is particularly useful in contexts where we are interested not merely in the maximal value of a function but in the specific input(s) that yield this maximal value.

I shall call this the *shallow consequentialist wrongness predicate*. We can contrast it with the *deep consequentialist wrongness predicate* W_d. For this, we define:

$$\bar{\chi}^{\mathsf{Val}}_{\textsc{MOCOR},D,C}(\mathsf{Val}(\mathsf{Out}_{D,C}(\phi))) \\ \text{if and only if} \quad \exists\, \psi \in \Phi_D \,:\, \mathsf{Val}(\mathsf{Out}_{D,C}(\psi)) > \mathsf{Val}(\mathsf{Out}_{D,C}(\phi))$$

Based on this 'anti' version of $\chi^{\mathsf{Val}}_{\textsc{MOCOR},D,C}$ we can then define

$$D,C \models_T W_d\phi \quad \text{if and only if} \quad \bar{\chi}^{\mathsf{Val}}_{T,D,C}(\mathsf{Val}(\mathsf{Out}_C(\phi))).$$

As long as we have a total order over the consequences (via their values) for a decision situation, W_s and W_d are obviously extensionally equivalent. Because then

$$\bar{\chi}^{\mathsf{Val}}_{T,D,C}(\mathsf{Val}(\mathsf{Out}_C(\phi))) \quad \text{if and only if} \quad \neg\chi^{\mathsf{Val}}_{T,D,C}(\mathsf{Val}(\mathsf{Out}_C(\phi))),$$

because

$$\exists\, \psi \in \Phi_D \,:\, \mathsf{Val}(\mathsf{Out}_{D,C}(\psi)) > \mathsf{Val}(\mathsf{Out}_{D,C}(\phi)) \\ \text{if and only if} \\ \neg\forall\, \psi \in \Phi_D \,:\, \mathsf{Val}(\mathsf{Out}_{D,C}(\psi)) \leq \mathsf{Val}(\mathsf{Out}_{D,C}(\phi)).$$

However, we will later encounter situations where

$$\mathsf{Val}(\mathsf{Out}_{D,C}(\psi)) \not\leq \mathsf{Val}(\mathsf{Out}_{D,C}(\phi))$$

does *not* entail

$$\mathsf{Val}(\mathsf{Out}_{D,C}(\psi)) > \mathsf{Val}(\mathsf{Out}_{D,C}(\phi))$$

because sometimes the values of the outcomes of the options ϕ and ψ are *incommensurable*. In these situations, W_s and W_d will disagree. However, for now, we can leave such formal details and can, finally, turn to the reconstruction of the CHALLENGE.

Chapter 3

The Challenge

This chapter is primarily reconstructive in nature. The principal objective is to anchor my understanding of the Challenge within the context of existing literature. This effort is twofold in its purpose. First, it ascertains that my overall project is grounded in a robust understanding of the subject, thereby minimizing certain dangers, most notably that of attacking a straw man. Second, engaging with the literature is advantageous as it brings to light various conceptual distinctions and allows accumulating classic instances of purported Troublemakers. Through the reconstruction of various interpretations of the Challenge and demarcating its scope as addressed in this work, this chapter establishes the foundation that is needed to tackle the Challenge systematically.

This endeavor has proven to be somewhat more laborious than initially anticipated. I believe it has been worth the effort, but I will preface it by explaining how and why I have chosen my approach and this specific corpus of literature.

3.1 On Choosing Giants

One of the goals of the first part of my thesis is to find relevant accounts of the Challenge anchored in the existing literature and, accordingly, to reconstruct plausible versions of valid arguments. Another goal is to distinguish the Challenge as addressed in my project from other, related challenges. Finally, I shall cast doubt on the idea that previous approaches have already solved the Challenge satisfactorily for camp moac. All of these endeavors are reconstructive in nature and require detailed engagement with a vast body of literature.

Such work is admittedly arduous and exhausting. Nevertheless, it is also crucial—for this particular project and for the progress of science as a whole. If one wants to stand on the shoulders of giants, one should first carefully determine *which* giants' shoulders one actually wants to stand on and how to best climb up there. The good thing is that one can try several giants and

different paths in different orders to find the best one. Accordingly, I avoid taking the reader up my own nearly decade-long winding path. Instead, I propose a route up, carefully mapped out at the cost of much blood, sweat, and tears. Several years of preliminary work have allowed me to add one or two climbing aids along the way. Hopefully, the result is an enjoyable, if non-trivial, path to new knowledge.

Given that the CHALLENGE has been extensively discussed in the literature over several decades—especially, though not exclusively, within the consequentialist camp—my goal carries a common risk of reconstructive efforts: the challenge of purposefully navigating a vast and complex body of work spanning generations. This literature often includes contributions from authors who may not have consistently acknowledged, much less referenced, each other's insights. At first glance, I am thus confronted with an immense and seemingly chaotic array of relevant works.

To bring the right kind of order to chaos, one has to define an epistemic goal. Depending on what one wants to achieve, some strategies are more appropriate than others. Since the present project is *not* an undertaking in the history of ideas, a structural pre-sorting seems promising in order to prepare a systematic analysis of the CHALLENGE. Accordingly, I will first group the relevant contributions along the specific *variant* of the CHALLENGE they consider.

The two main clusters of work correspond to distinct strands in the literature, each framing the CHALLENGE as a genuinely (or at least primarily) consequentialist challenge, yet approaching it from fundamentally different perspectives. These strands, in a sense, complement each other in ways that deserve and remain to be uncovered. The first cluster frames the CHALLENGE as one of intra-theoretical inconsistency. The second cluster sees it rather as a violation of (consequentialist) intuitions. While the first formulation of the CHALLENGE is based on more presuppositions, it would also, if successful, be all the more fatal for camp MOAC. Together, the two clusters represent the most relevant contributions in terms of my specific research question.

A different way of dividing the contributions that will occupy us in this chapter provides us with an important distinction that enables a systematic treatment of the CHALLENGE along another dimension. Apart from minor stylistic differences, all authors discuss cases with the same general basic properties—Individual Optimality and Collective Suboptimality—that they claim to give rise to the CHALLENGE. I have given these cases the name Troublemakers in the introduction. Nevertheless, one can and should distinguish between several different *types* of Troublemakers. These types differ with respect to the underlying structure that gives them the relevant properties.

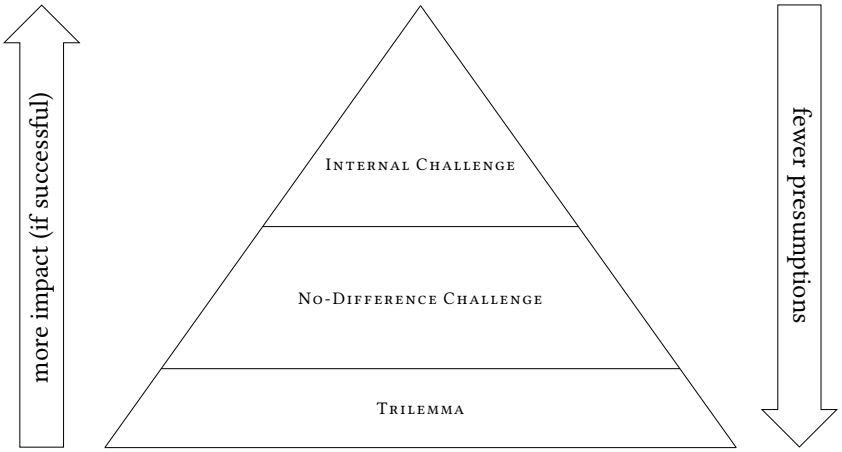


Figure 3.1: The three variants that make up the Challenge in terms of the Pyramid as tackled in this thesis, ordered by their strength and the extent of their preconditions.

Furthermore, while some contributions focus only on one specific kind of Troublemaker, others discuss several types. Thus, it seems that to approach the Challenge systematically, one should have a clear and precise taxonomy of Troublemakers. Along the way, I use the opportunity to collect various specimens in this chapter and sketch a taxonomy that will be refined and formalized, at least a little, in the context of the second part of my project. Furthermore, along the way, I distinguish the Challenge from two related challenges.

The number of contributions that concern the Challenge in some way is far too large to be treated exhaustively here. Instead, it makes sense to focus on particularly influential contributions, i.e., pieces that are particularly deeply connected and well-connected within the debate, and on contributions that help us draw useful conceptual distinctions—even though this might not do justice to all valuable contributions out there. Therefore, I must decide which works will be examined and to what extent. All this work, then, will ensure that my project does not burn down a straw man.

3.2 The Pyramid

This project aims to defend moac, and thus a very specific family of moral theories, against the Challenge. In the introduction, I sketched the Challenge in what could be considered its most menacing form for moac, which I called Internal Challenge. That particular form, however, is underpinned by quite a number of theoretical presuppositions. As a result, this variant of the Challenge presents a relatively extensive "surface for (relief) attacks" for proponents of moac. As previously indicated, however, other variants of the Challenge involve fewer presuppositions but also come with less 'punch', i.e., have less dire consequences for moac if successful.

One question that emerges then is which variant should be subjected to scrutiny. In line with the principle of charity, it is essential to engage with the most promising version of the challenge at hand. This raises the question of what the Challenge's most formidable version is.

The correct answer, I believe, is actually an evasion of the question. The strongest form of the Challenge is not a *single* variant but rather a hierarchically ordered *sequence* of variants, each backing up those above it in the hierarchy. Only once it is shown that moac is immune to every variant can the Challenge be considered resolved. I call this hierarchical structure of variants of the Challenge the Pyramid (cf. Figure 3.1).

The Pyramid consists first of the Internal Challenge, already outlined in the introduction (and about to be reconstructed with more care in Section 3.5), which has haunted objective consequentialists for decades. In the second row lurks the Challenge in the form of a No-Difference Challenge, a formulation that has yet to be reconstructed (this will be done in Section 3.4) and is also quite prominent in the literature. The last row presents a very general formulation of the Challenge, one that is not specific to any particular theory and is intended to challenge an entire set of moral theories, not just consequentialist ones. Due to the chosen representation, which is to be introduced in the following Section 3.3, I call this variant simply the Trilemma.

With each additional layer of the Pyramid, new theoretical assumptions are required to formulate the respective version of the Challenge. On the other hand, the versions represented by the lower levels are based on more assumptions presumed to be 'intuitively convincing'. The rest of this chapter is devoted to reconstructing the layers of the Pyramid, starting from the foundation and working our way to the top.

3.3 The Trilemma

Before delving into the two moac-centric variants of the Challenge, which correspond to what I identify as two separate strands within the relevant literature, I will discuss a theory-agnostic variant underpinned by a minimal set of presuppositions. This initial variant is exceptionally well suited to serve as our point of entry because it predominantly relies on intuitive considerations, allowing us to sidestep, for the time being, the intricate and sometimes nebulous terrain of moral philosophy and, more specifically, of consequentialist theorizing.

The variant, which draws inspiration from David Estlund's formulation (see Estlund 2017, pp. 53–55) within a non-consequentialist framework,[42]

[42]Estlund employs the Trilemma as one component of a companion-in-guilt argument to bolster his particular, collective notion of justice. We can ignore the details of that notion,

encapsulates the fundamental essence of the Challenge through the lens of a trilemma, which, in the interest of brevity, I will refer to simply as the Trilemma.

To formulate it, we first need a case with the 'right' structure. Here is a specific case,[43] borrowed from Felix Pinkert (2015, pp. 973–975), that will serve as the standard example throughout most of this book:

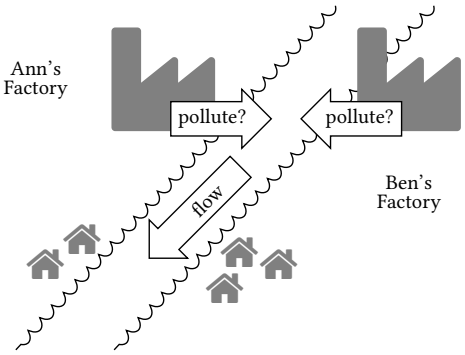


Figure 3.2: The Two Factories case.

Case 3.1 (Two Factories) *Ann and Ben each own a factory along the same river. Both can choose to produce either cleanly or cheaply, thereby polluting the river. The local market is highly competitive. Thus, a factory that produces cleanly would become noncompetitive if the other factory pollutes. The local social system is underdeveloped, and the economic situation is terrible. Hence, if a factory closes, this will cause significant unemployment and social hardship. If at least one factory produces cheaply, the resulting pollution will eventually destroy the local ecosystem and erode the livelihood of a village downstream. However, any additional polluter would not make the situation worse in this regard. Ann and Ben decide and act independently in the sense that neither of them can coordinate with the other, nor can any agent observe the actions of the other. Thus, whatever each agent does, they would do it regardless of what the other does.*

Two Factories involves several agents who are in a certain kind of interdependency, which obviously deserves to be investigated in more detail in the course of this chapter. For now, we can note that, to all appearances, this dependency concerns the outcomes (and not the decision-making of the

which grapples with a challenge analogous to the one faced here, but it is important to note that Estlund's Trilemma is formulated in terms of moral obligations rather than moral rightness. Consequently, the Trilemma discussed here diverges significantly from Estlund's original version and should, despite being heavily influenced by it, not be perceived as semantically equivalent.

[43] Estlund offers a structurally identical example involving two doctors, Dr. Stitch and Dr. Patch (cf. Estlund 2017, pp. 53–55), who together could save a patient's life, though neither can achieve this alone. In my view, however, Estlund's example is less suited to illustrating this very general variant of the Challenge, as it tends to evoke intuitions about special obligations stemming from the fact that both agents are medical professionals—intuitions that can detract from the main issue. Both Pinkert's and Estlund's examples will be introduced and discussed in detail later. These cases all belong to a class sometimes referred to as "High-Low Cases", a term that Christopher Woodard partly attributes to Michael Bacharach (cf. Woodard 2003, p. 217).

agents, which is explicitly *in*dependent), and it prevents the decision situations from being considered in isolation in a trivial way, i.e., we cannot simply decompose the situation into two individual decision situations, one for Ann and one for Ben. It therefore seems appropriate to understand Two Factories as a description of a *collective* decision situations rather than 'just' two individual ones. The relevant form of dependency is shown mainly by the fact that, whatever the final outcome of this situation might be, it does not depend on the action of any single agent, but depends on the specific *combination* of actions that is carried out. Here is a tentative definition of such collective decision situations:

Definition 3.1 (Collective Decision Situation – tentative)
A collective decision situation *is a situation in which multiple agents are each presented with multiple options, and within a given context, each combination of actions has an associated consequence.*

It will become apparent in the course of this chapter that Definition 3.1 does not quite do justice to the actual complexity of collective decision situations. However, a more detailed, precise, and even formal elaboration of collective decision situations can be reserved for the second part of my project. For now, Definition 3.1 is sufficient, although it glosses over some issues. Most importantly, we will just have to implicitly assume in the following—much in line with the debate we explore in this chapter—that there is a way to reason about what is the right thing to do for agents in such collective decision situations. During the course of this chapter, the common way of consequentialist reasoning in collective situation situations will become more and more clear.

Back to the Challenge. The Challenge in its trilemmatic variant arises as soon as we assume that Ann and Ben both pollute, which apparently constitutes a troublesome combination in the sense defined in Definition 1.1. Recall:

Definition 1.1 (Troublemakers – tentative) *A collective decision situation is a Troublemaker if and only if there is a troublesome combination of options therein, i.e., the agents can act in ways such that*

> **(Collective Suboptimality)** *together they would produce a morally suboptimal outcome and*
>
> **(Individual Optimality)** *none of them could make a difference for the morally better by unilaterally acting differently.*

The Trilemma now consists of the following three statements, all of which appear to be *prima facie* true given the instantiation of that troublesome combination. For example, let us assume that both Ann and Ben actually opt for polluting:

(H_1) Something wrong happens if the ecosystem of the river and the livelihood of the adjacent village are destroyed.

(H_2) If something wrong *happens*, then because someone *did* wrong.

(H_3) No one did wrong.

Here are a few *prima facie* good reasons to believe that all three statements are true. First, given that both Ann and Ben pollute, a lot of harm is done: the river's ecosystem is destroyed, and the livelihood of the nearby village is eroded. All this is not the result of some calamity of nature or the consequence of a chain of unfortunate circumstances, but result of *actions*. At the same time, a much better outcome was available through another course of action. If Ann and Ben had both produced cleanly, they would have brought about a situation with all of the benefits of the current one but none of the harm. This clearly would have been better. Therefore, H_1 seems intuitively convincing: The situation that is brought about if both Ann and Ben pollute is a morally unbearable result of actions; something has gone wrong morally.

Second, H_2 seems convincing on conceptual grounds alone. If something is wrong and not merely lamentable, it is because it is a consequence of an *action* (see also Principle 2.1 in Section 2.2). A consequence, moreover, of at least one *wrong* action—which must then have been performed by someone, as there can be *no action without agent*. As David Estlund (Apr. 27, 2017) puts it: "If something is morally wrong, then there was an obligation on some agent to act or omit other than as they did." Even if one should find these conceptual considerations questionable,[44] we can simply state the logical consequence of H_1 and H_2 directly. Together, they imply that, in the present case, someone did wrong.[45] Instead of inferring this from H_1 and H_2, we can simply reduce the two to

($H_{1,2}$) Some agent did wrong.

In a sense, this is what Frank Jackson does (Jackson 1987, p. 100) when he judges—for a structurally identical case—that it "is evident that something wrong happens [...]; but more than that is evident: something wrong is done." If we reduce H_1 and H_2 in this way, the TRILEMMA collapses into a very simple dilemma.

[44] It may justifiably be argued that in the case of Estlund's formulation in terms of obligations, the matter is even clearer than in the case of 'mere' moral wrongdoing. But this again depends on so many theoretical presuppositions (cf. Section 2.2) that we need not worry about it here and can put this concern aside for the sake of argument.

[45] Be aware that [p because q] implies not only that p but also that q (at least in general—I will later argue that this apparent truism should be taken with a grain of salt).

Finally, given Ben's actual actions, the river will be polluted regardless of what Ann does. This makes it difficult to explain why Ann's act of polluting constitutes wrongdoing. The same applies in the reverse case: Given that Ann actually pollutes, it becomes equally challenging to explain why Ben's act of polluting is an act of wrongdoing, as the river will be polluted regardless of his actions. The only effect of *not* polluting would be additional harm through unemployment, worsening the situation even further. In other words, the following conditionals are warranted (against the background of Two Factories plus the fact that both agents polluted the environment):

(4) If Ann had produced cleanly, nothing would have been better, but some things would have been worse because her workers would have lost their jobs.

(5) If Ben had produced cleanly, nothing would have been better, but some things would have been worse because his workers would have lost their jobs.

Thus, they both *did the best they could have done given what the other one did*. None of them could have made a difference for the better by (individually) doing otherwise. These considerations strongly support H_3.

As is the nature of trilemmas, not all of these propositions *can* be true, i.e., this triad of propositions is inconsistent: H_3 denies precisely what follows from H_1 and H_2 (or *vice versa*). Which one should go? This is the CHALLENGE in pre-theoretic form.

I assume that few would contest H_1,[46] while most disagree with H_3. (H_2 will probably mainly elicit strong opinions from ethicists, who I think will mostly agree.) H_3 seems to be indeed somehow off. Michael J. Zimmerman (1996, p. 257) perhaps put the finger on the relevant core intuition when he wrote that it seems as if in such cases there "is a sense in which [...] two wrongs [...] make a right." The matter would be straightforward if we think about the individual case (recall Factory). Polluting would be wrong. But when two such acts come together, they seem to oddly 'exculpate' each other.

There are undoubtedly many promising avenues for rejecting H_3 on systematic, principle-based grounds, with various moral theories supporting different approaches: Ann and Ben might neglect certain *duties*, such as the duty

[46] It must be pointed out that we are on pre-theoretical ground here. *De facto*, I am well aware of some philosophers who would not accept certain involved formulations. We should not be bothered by this at this point. The only aspect that matters here is that the TRILEMMA corresponds to a (rarely explicated, see especially Estlund 2017) variant of the CHALLENGE that comes up every now and then in the discourse. It does not have a particularly supporting role in this project but forms the final line of defense of the camp CHALLENGE and serves here primarily as a starter, intuition trigger, and pre-theoretic motivational basis.

to preserve the environment or to do their part and to keep their hands clean, they might infringe upon the *rights* of others through their polluting actions; or their behavior might exhibit particular *vices*. But which one of these approaches is justified? That depends on what moral theory is correct—and it doesn't look like camp MOAC is in a promising position. Thus, if we want to resolve the TRILEMMA, we need a theoretical framework that allows for a deeper analysis. We must leave the cozy realm of pre-theoretical reasoning.

The above considerations apparently in favor of H_3 are essentially based on the observation that no agent could have changed anything for the better by acting differently, cf. (4) and (5). Therefore, the problem of refuting H_3 and thus the resolution of the TRILEMMA via the *prima facie* most plausible move, is particularly difficult for theories which we shall call Difference-Making Views. Such views are characterized by the fact that they accept the Difference Principle. Here is Frank Jackson's (1987, p. 94) formulation,[47] which will serve as a starting point:

Principle 3.1 (Difference Principle) *The morality of an action depends on the difference it makes; [i.e.,] it depends on the relationship between what would be the case were the act performed and what would be the case were the act not performed.*

Jackson insists that his formulation is attractive for a broad class of moral theories. First, because the principle "as stated says nothing about how to evaluate the differences, and nothing about what kinds of differences matter morally" (ibid., p. 94).[48] Second, because it allows that the morality of an action depends *not solely* on the difference it makes, but also on other grounds like, say, intentions or duties. It just states that the moral status of an action (or its "morality", in Jackson's terms) is *also* dependent on (or a function of) the differences it makes.

Whether this is true or not, it can be said that H_3 in particular should be difficult to reject for Difference-Making Views. Recall (4) and (5) above. If they were indeed both true, it would be hard to see how any plausible Difference-Making View – especially one that is maximizing in the sense that MOAC desires to be—could identify any wrongdoing in Two Factories (given that the agents actually did pollute). We recall that MOAC theories were characterized by their criterion of rightness, MOCOR:

[47] Jackson's formulation raises several questions, perhaps the most important for this project being which states of affairs are meant by "what would be the case were the act not performed". These and related questions will be set aside for now to focus on the reconstructive effort in this chapter. I will address these issues in greater detail in the second part of this project. For the moment, we can assume that these questions are primarily about articulating this relationship more precisely and that a reasonable way to do so *indeed* exists.

[48] The connection to the debate on the 'consequentializability' of moral theories should be evident; see, for instance, Dreier 2011; Portmore 2007, 2009.

Criterion 1.1 (MOCOR – tentative) *An action is right if and only if there is no alternative action that would lead to better consequences.*

Since the consequences of an action are precisely those differences that the action makes in the sense expressed in the Difference Principle, MOAC theories, of course, *are* Difference-Making Views. After all, any objective (act-)consequentialist theory can be characterized by an 'enhanced' version of the Difference Principle:

Principle 3.2 (Consequentialist's Creed) *The morality of an action depends* solely *on the difference it makes; i.e., it depends on the relationship between what would be the case were the act performed and what would be the case were the act not performed,* and on nothing else.

Accordingly, the Trilemma is particularly hard to solve for MOAC – at least if its supporters really wanted to attack H_3.

The question, then, is how impactful the potential punch of the Trilemma truly is. Maybe it just shows us something interesting about the non-consequentialist intuitions underlying the agreement with H_1 and H_2. It could be, for example, that consequentialists make the following considerations, which may undermine H_2: H_1 merely expresses that there would have been a better possible outcome in the sense that it could have been produced by both agents together (cf. Collective Suboptimality in Definition 1.1). Then we should rather say that something *needlessly bad* happened. This is perfectly compatible, as Two Factories shows, with the observation that there are cases where, given what the agents actually do, no individual agent could have produced a better outcome (cf. Individual Optimality in Definition 1.1). Then, consequentialists might infer that it does *not at all* follow from the occurrence of something needlessly bad that there must be a wrongful action by any agent to account for that badness (or wrongness).

But perhaps MOAC's inability to reject the Trilemma shows more. It might show that MOAC theories cannot do justice to certain basic *consequentialist* intuitions. This leads us to the formulation of the Challenge as the No-Difference Challenge, which is consistent with the first, weaker strand found in the literature. Alternatively, this inability of MOAC could indicate an even more profound failure in the sense of the Internal Challenge outlined in the introduction, as claimed by the contributions to the second strand. Both traditions of framing the Challenge—the two main variants of the Challenge—areexplored below.

3.4 The Challenge as No-Difference Challenge

According to the first, *weak* strand (which is weak in that it involves fewer assumptions than the Internal Challenge sketched in the introduction), the Challenge arises when certain actions are *intuitively* wrong (right), but the intuitively right (wrong) alternative action, presumably, would make (or would have made) no difference for the better (worse). Looking back to the Trilemma and Two Factories, the Challenge would then be that it is assumed that H_3 has to be rejected—but at the same time, one cannot do so because of the apparent individual inefficacy. Call this apparent inability of moac to come to *intuitively* adequate assessments in Troublemakers the No-Difference Challenge.

In other words, the No-Difference Challenge locates the Challenge in the apparent commitment of consequentialist views to *counterintuitive* moral assessments, given the apparent inability of individual agents to make a difference in certain cases. Thus, these allegedly violated intuitions are explicitly 'consequentialist in spirit'. Here is a quite recent formulation by Holly Lawford-Smith and William Tuckwell (2020, p. 635):

> It is an objection to act consequentialist views (but also other difference-making views) that there are classes of actions that don't make a difference and yet we seem to have a strong intuition that those actions should not be performed. Our intuitions suggest that these actions are wrong; the argument from no difference (including insignificant difference) is that they're not wrong because they make no difference. Cases where the actions of many different people add up to cause harm at the level of the collective are prominent examples of where the No-Difference Challenge arises [...]

The No-Difference Challenge constitutes an essential part of the general debate about the Challenge, which by itself is reason enough to study it in the context of my project. But even apart from this, engaging with the No-Difference Challenge debate proves to be a fruitful endeavor in several ways. In the following, we collect some relevant observations.

First, the No-Difference Challenge is *broader* than the Challenge. This is partly due to the fact that the set of cases that raise the No-Difference Challenge is broader than the set of Troublemakers. Recall, again, the tentative definition:

Definition 1.1 (Troublemakers – tentative) *A collective decision situation is a Troublemaker if and only if there is a troublesome combination of options therein, i.e., the agents can act in ways such that*

> **(Collective Suboptimality)** *together they would produce a morally suboptimal outcome and*
>
> **(Individual Optimality)** *none of them could make a difference for the morally better by unilaterally acting differently.*

We can put the characterization from the preceding quotation into a form[49] similar to that definition:

Definition 3.2 (No-Difference Case (tentative))
A situation is a No-Difference Case if and only if there is at least one agent that can act in a way such that

> **(Intuitive Wrongness)** *it is intuitively morally wrong, but*
>
> **(Individual Optimality)** *the agent could not make a difference for the morally better by unilaterally acting differently.*

Thus, as also suggested by the last sentence of the above quotation from Lawford-Smith and Tuckwell, the No-Difference Challenge can, in principle, arise in the absence of collective contexts.[50]

Conversely, the definition of Troublemakers does not imply that the actions involved are intuitively wrong, but only that the actions which together constitute a troublesome combination lead to *suboptimal* outcomes. Since Definition 1.1 does not mention intuitive wrongness at all but rather Collective Suboptimality, we should assume that Intuitive Wrongness is satisfied in some Troublemakers (for at least one involved action), but not necessarily in all.

Therefore, the set of No-Difference Cases is distinct from that of the Troublemakers, even if the Challenge is *understood as* a No-Difference Challenge. Thus, that some case is a Troublemakers does not entail that it is a No-Difference Case and vice versa. Nevertheless, the No-Difference

[49] A terminological side note is necessary: I call this class of cases No-Difference Cases simply because it fits with the established name of the challenge they raise, i.e., the No-Difference Challenge. But we will see below that this name can quickly become misleading. For example, one might be tempted to think primarily, or even exclusively, of cases such as the climate change scenario described in the introduction, in which, by assumption, each individual action makes no morally relevant difference to some morally catastrophic whole to which individually insignificant contributions accumulate. Those cases, which we later call Cumulative Effects Cases, are included here but not exclusively addressed. More on this below.

[50] For example, Warren Quinn's puzzle of the self-torturer can be understood not only as one of instrumental rationality but also as a moral problem in the sense of the No-Difference Challenge (cf. Quinn 1990). Those who see the self-other asymmetry (cf. Slote 1984) as a stumbling block for a moral re-framing of the puzzle may think of a modified version in which one's actions affect another person and not oneself.

Challenge is *commonly* formulated and discussed in the context of collective decision situations, i.e., it is usually about collective No-Difference Cases (see also Figure 3.3). Whenever I refer to Troublemakers in the context of the No-Difference Challenge, I refer to collective No-Difference Cases if not explicitly stated otherwise.

Second, the No-Difference Challenge is broader in another sense as well. It has a plausible variant that also concerns other varieties of act-consequentialist theories, even if we restrict ourselves to collective contexts. In particular, some of the more recent contributions to the debate (Hedden 2020; Kagan 2011) are examples of a collective No-Difference Challenge for subjective moral theories. As we will see, this cannot be said for the Internal Challenge. In this respect, a clearer understanding of the difference between the No-Difference Challenge and the Challenge helps to frame the present project more clearly and to demarcate it from other collective challenges of consequentialism which fall outside its scope.

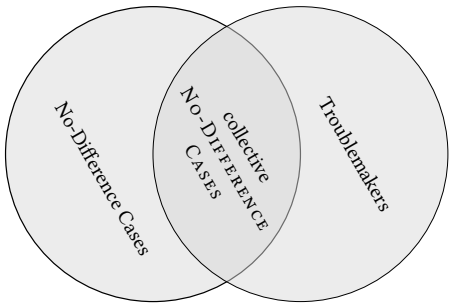


Figure 3.3: Neither are all Troublemakers No-Difference Cases, nor the other way around. For this project, only those No-Difference-Cases are of interest that are No-Difference Cases due to collective contexts, i.e., that are not only No-Difference Cases but also Troublemakers. We can call these cases in the intersection of the two sets "collective No--Difference Cases". It's stipulated that in this intersection are only cases in which the intuitive moral failure is accompanied by collective suboptimality.

Third, the Challenge understood in terms of[51] the No-Difference Challenge represents a kind of fallback for the Internal Challenge because it works with relatively few presuppositions and, thus, comes with less theoretical baggage than the Internal Challenge. We only need some intuition-based judgments.

Fourth and finally, the discussion of central contributions to the No-Difference Challenge also allows us to make some relevant conceptual and theoretical distinctions and to introduce examples that will prove useful in the sequel.

It is a good idea to approach the No-Difference Challenge by starting with the contribution that some claim has "introduced [it] to philosophers" (Lawford-Smith and Tuckwell 2020, p. 634), namely with Johnathan

[51] Hereafter, reference to the No-Difference Challenge is always to be understood as referring to the Challenge as, or in terms of, the No-Difference Challenge—should I wish to refer to some instance of the No-Difference Challenge that is devoid of collective context, I will make it explicit.

Glover's article "It Makes No Difference Whether or Not I Do It" (cf. Glover and Scott-Taggart 1975).[52] Even though this might be a slight overstatement,[53] it is fair to say that Glover's contribution remains extremely influential. Accordingly, Glover's account also is a good starting point for the reconstruction of the Challenge as a No-Difference Challenge.

Glover begins with an observation regarding what he calls "a family of arguments relating to the insignificance of a single person's act or omission" (ibid., p. 172). He offers two typical statements exemplifying the practice:

(6) If I don't do it, someone else will.

and

(7) One person makes no difference.

Glover refers to such arguments as "attempted justifications", and the central question of the No-Difference Challenge is whether they truly justify the acts in question. While Glover emphasizes that, in many instances, such statements are simply false, he acknowledges that in other cases, the matter is far from clear. In such situations, the inability to make a difference for the better appears to obstruct an intuitively correct moral judgment. In other words, the No-Difference Challenge, like Lawford-Smith and Tuckwell insisted above, indeed presupposes at least the Difference Principle. (Furthermore, Glover explicitly refers to consequentialism, even though he leaves unclear what exact variant of the theory he has in mind and whether he restricts his considerations to consequentialist theories.)

Let us return to the two statements and their relation to each other. At first glance, one might wonder whether (7) isn't just referring to a broader class of cases than (6) is, i.e., whether (7) isn't entailed by (6). This, however, is not

[52] Strictly speaking, this piece has two authors, but it is a bipartite article, and the second part, i.e., the one by Scott-Taggart, is not of interest to this project as he focuses on mere excuses and individual responsibility in collective contexts and not on actual right-doing.

[53] As we will see later, a version of the No-Difference Challenge can already be found in a piece by C. D. Broad (1916; see also Subsection 3.5.1). For the specific version of the No-Difference Challenge relevant in the context of this project, the statement by Lawford-Smith and Tuckwell is arguably correct, though. However, No-Difference Cases have been discussed, both explicitly and implicitly, earlier in the context of act-consequentialism (cf. Smart and Bernard Williams 1973, discussed in some detail in Subsubsection 7.3.2.3 and Subsection 8.3.1). In particular, the literature on rule-consequentialism is replete with examples of such cases. For instance, R. F. Harrod argued, early in the corresponding debate, that there "are certain acts which, when performed on *n* similar occasions, have consequences more than *n* times as great as those resulting from one performance" (cf. Harrod 1936, p. 148). Other important examples are given by Gibbard 1965 and Brandt 1959. We will revisit both later, albeit briefly, as they have inspired J.J.C. Smart and Donald Regan, respectively (cf. Regan 1980, which is discussed in more detail later in this chapter).

what Glover had in mind. Properly understood, (6) and (7) cite two distinct reasons for believing in the alleged individual inefficacy or, put differently, refer to fundamentally different structures that give rise to it. Furthermore, only (6) involves an explicitly collective context. (7), on the other hand, merely refers to the absence of a difference in general, without saying *why* no relevant difference is being made. It gets by without reference to other agents or other actions. Glover, thus, takes the two statements to represent two distinct *types* of collective No-Difference Cases (and of Troublemakers).

Two concrete examples help us understand the distinction that Glover had in mind. Concerning (6), Glover cites an example from Bernard Williams (cf. Bernard Williams 1973, p. 124, modernized a bit by me in the following):

Case 3.2 (Job Market) *George, a family man with two children who has just taken his PhD in chemistry, finds it extremely difficult to get a job. He is not very robust in health, which cuts down the number of jobs he might be able to do satisfactorily. The current situation makes him and his family significantly worse off than if George had a well-paid job. An older chemist, who knows about George's situation, says that he can get George a well-paid job in a laboratory, which pursues research into chemical and biological warfare. George says he cannot accept this since he is opposed to chemical and biological warfare. The older man replies that he is not too keen on it himself, come to that, but after all, George's refusal is not going to make the job or the laboratory go away; what is more, he happens to know that if George refuses the job, it will certainly be offered to a contemporary of George's, Paul, who is not inhibited by any such scruples. If Paul were to get the job, he would push along the research with greater zeal than George would. Indeed, it is not merely a concern for George and his family, but (to speak frankly and in confidence) some alarm about this other man's excess of zeal, which has led the older man to offer his influence to get George the job.*

If George took the job, he would invest his time and labor into the research and development of chemical and biological weapons, which we might, plausibly and for the sake of argument, understand as an unacceptably bad outcome. However, if George were not to take the job, the much more strongly motivated Paul would take it instead and would put significantly more effort into this harmful endeavor. The best outcome would obviously be that the job remains vacant.

We can visualize the situation by representing Job Market in an *extensive form*, a tree-like, graph-based structure, as illustrated in Figure 3.4. Extensive forms explicitly and visually encode the order of actions: George has to decide first; then, if he rejects the offer, Paul has to make his decision. Let us call such collective decision situations where, by description, the order of action is defined (and crucial) Sequential Cases.

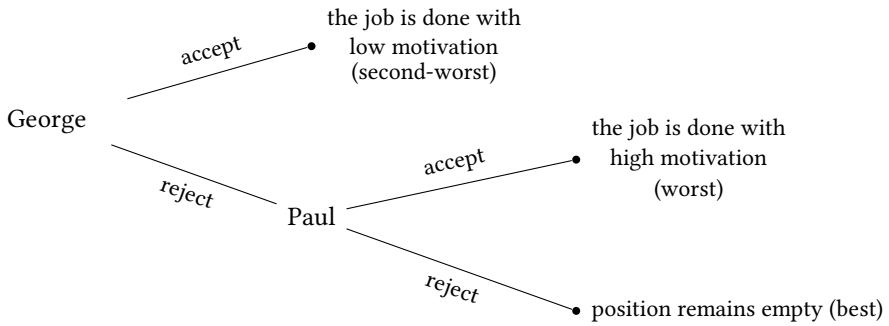


Figure 3.4: The extensive form of Job Market.

Strictly speaking, Job Market qualifies as a No-Difference Case (and as a Troublemaker) *only* if we are willing to extend our understanding of combinations of actions to include hypothetical (and potentially counterfactual) actions, such as Paul's acceptance of the job if it were offered to him. Only then does the combination of George accepting the job offer, together with Paul counterfactually taking the job in the scenario where George declines, constitute a troublesome combination: George's acceptance leads to an outcome that is intuitively unacceptably bad and certainly suboptimal. After all, in principle, a better outcome could have been brought about, namely, if both George and Paul refused the offer and the position remained open. Let's agree that it is intuitively wrong for George to accept such an offer. However, it is not under George's control to bring about the only better outcome. If George did *not* take the job, the much more motivated Paul would be offered it instead—and Paul would take it, with worse results. Therefore, to all appearances, Job Market is a collective No-Difference Case (and a Troublemaker), and George can utter (6) truthfully. However, George cannot express (7) truthfully. For although he cannot make a difference for the better, his action *does* make a difference. After all, it would be worse if he didn't take the job because then Paul would step in.

We can, of course, paraphrase Job Market such that we can do without hypothetical actions as components of troublesome combinations. For this, we can imagine that the chemical weapons company only needs one employee to pursue its harmful and damnable plans, but that they have taken the precaution of advertising *two* jobs because there is more than enough money to be made from chemical weapons anyway and in times of a shortage of skilled workers, you take every chemist you can get. Assuming that both of them accept, George could still have a moderating effect on Paul's overzealousness for evil; if Paul alone accepted, he could do as he pleased, which would have worse consequences; if George alone accepted, he would still contribute something to the bad cause, but much less than they would together. Last but not least, it would be best if the position remained vacant.

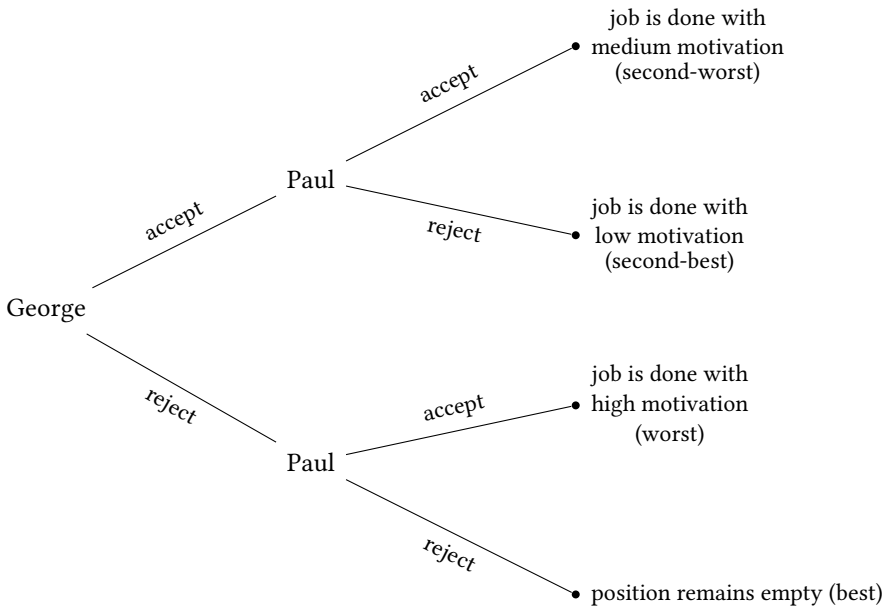


Figure 3.5: The extensive form of the modified Job Market case.

Because of the good relations with George's older colleague, they offer him the job first. This modified version of Job Market corresponds to the extensive form shown in Figure 3.5. Other examples with very similar structures are cases of two killers shooting the same person or two perpetrators poisoning a victim and then shooting him before the poison takes effect (Jackson 1987, 1997; Parfit 1984; Zamir 2001).

All these cases are *minimal* in the sense that they involve two agents, each with two possible courses of action. With fewer agents, the collective aspect would be lost; with fewer options, there would be no decisions to make. Nevertheless, larger cases can also belong to the category of Sequential Cases. Admittedly, larger cases are challenging to grasp, particularly those involving numerous agents.[54] The most famous such case is Kagan's chicken example (cf. Kagan 2011, heavily modified below for the sake of simplicity):

Case 3.3 (Chicken Counter) *Shelly is thinking about becoming a vegetarian. He usually buys exactly one chicken a week from his friend Tommi's local deli and doesn't eat any other meat at all. From various casual discussions with Tommi, he knows how the business works: From the local poultry farm, Tommi gets a certain number of chickens delivered every week, which roughly corresponds to the number of sold chickens the week before. The details are like this: There is a digital counter in his cash register that is set to 0 when a new delivery arrives. Whenever Tommi has sold 10 chickens, his cash register automatically*

[54] In the following, we briefly address non-minimal cases in terms of the number of agents, while disregarding cases with a greater number of options.

sends a notification to the local poultry farm. This notification causes them to breed (and thereby effectively torture) another 10 chickens. If no new orders come in, they do not breed (and thereby effectively torture) any new chickens.

Of course, this scenario is completely artificial and, in many ways, extremely simplistic. But it is enough for us to capture the essential intricacies of real consumer decisions in the form of a Troublemaker. For this, we need a few more assumptions, though:

First, we assume the following distribution with respect to the actual decisions made for or against chicken purchases this week:

Fact 3.1 *85 people considered buying a chicken at Tommi's local deli this week, 81 of which purchased a chicken.*

We also make two assumptions regarding the distribution of moral values:

Fact 3.2 *Each consumption of a chicken is accompanied by at least some positive moral value (e.g., pleasure).*

and

Fact 3.3 *The total pleasure individuals derive from eating chicken is outweighed by the total suffering endured by chickens during their upbringing.*

Figure 3.6 shows a tiny extract of the corresponding collective decision situations. Note that a node labeled $\langle x, y\rangle$ corresponds to the situation where y decisions have led to x chicken purchases so far. "$\langle x, y\rangle$" can thus be read as "x of y visitors have opted for a chicken purchase". In light of Fact 3.1, we may stipulate[55] that the uppermost path, ending in the state $\langle 81, 85\rangle$, corresponds to the *actual* combination of actions.

This is a troublesome combination of actions. First, it is collectively suboptimal: we can quickly identify a significant number of possible combinations of actions that lead to some better result. While the exact number and composition of these combinations depend on the concrete distribution of benefits and harms that is not specified in detail here, we can say for sure that any state $\langle x, y\rangle$ with $x \leq 9$ will lead to a better outcome. After all, according to Fact 3.2, the corresponding consequences would involve some pleasures of eating chicken but no additional harm, as no new chickens would be ordered (and thus raised). Hence, these corresponding consequences have an overall

[55] To be able to single out *the* path that corresponds to the actually performed combination of actions, we actually would need more information, namely which customer bought and which did not. Giving such a list would be a bit lengthy, and I spare us this and further complications. We may simply assume, without loss of generality, that said path corresponds to the actually performed combination of actions.

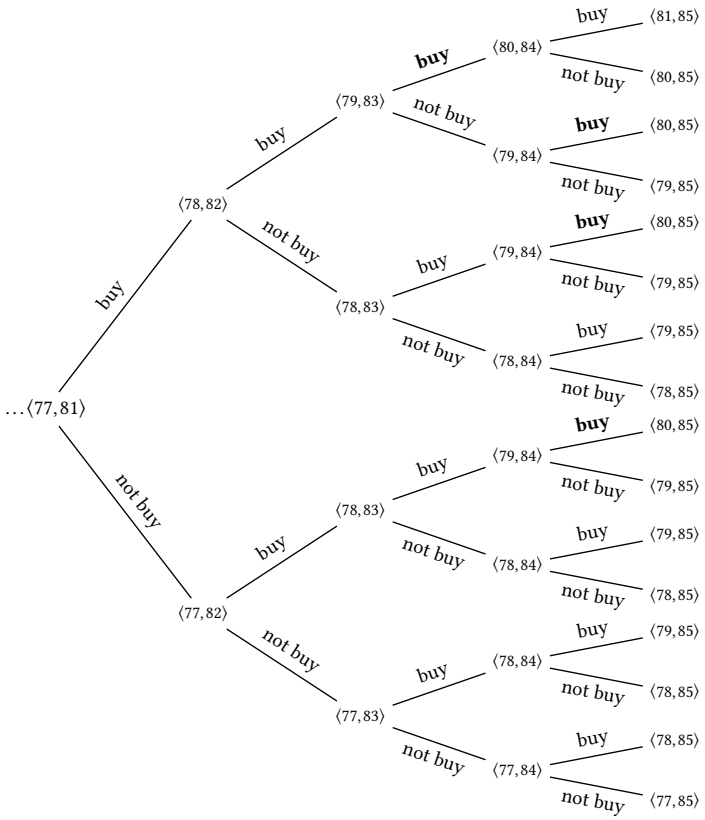


Figure 3.6: An excerpt from a possible manifestation of Chicken Counter for 85 consumer decisions (starting from the 82nd decision). A node labeled $\langle x, y\rangle$ represents a situation where y decisions have so far resulted in x chicken purchases. Buying decisions in bold indicate actions that triggered an order for 10 additional chickens.

positive value. To the contrary, Fact 3.3 warrants that the actually produced outcome, corresponding to a $\langle 81, 85\rangle$ state, is overall negative.

Second, the state $\langle 81, 85\rangle$ is individually optimal: For each and every involved agent, indeed, refraining from their chicken purchase would only cost them the corresponding amount of pleasure, but would not spare any chicken its cruel fate.[56] After all, if any individual agent had acted otherwise, 80 instead of 81 chickens would have been consumed, and, thus, the very same number of chickens would have been raised.[57]

[56] To be precise, forgoing chicken purchase costs them some pleasure only *comparatively*, i.e., insofar as it interferes with their chicken consumption (their source of value, by assumption, cf. Fact 3.2). We can easily make this further complication irrelevant by stipulating that there are no further sources of (dis-)value involved. (Suppose, for example, that some people are feeling bad for choosing *not* to buy chicken. Such affective reactions may sound implausible at first, but I think they can actually explain many a behavior we observe in reality.)

[57] Thus, any combination of actions ending in a state $\langle x, y\rangle$ with $x \bmod 10 = 0$ is *not* a troublesome combination of actions.

Chicken Counter makes it particularly clear that the Sequential Cases discussed so far, i.e., Sequential Cases that all Troublemakers, specifically involve certain triggers[58] or *thresholds*. These triggers are activated by some of the actions that constitute some troublesome combination, respectively; however, other actions *would* have taken the place of each action had it failed to occur. I shall call such cases Threshold Cases. Figure 3.7 shows the relevant relation of kinds of cases.

Threshold Cases are Troublemakers qua involving some kind of 'adverse causal dependency', but the underlying nature of these dependencies is importantly different. I will call structures like those in Job Market *preemption* and structures like in Chicken Counter (given an unfortunate number of actual purchases) *overdetermination*. Overdetermination and preemption are both concepts in the philosophy of causation, specifically addressing situations where multiple factors appear to be involved in causing a particular effect.

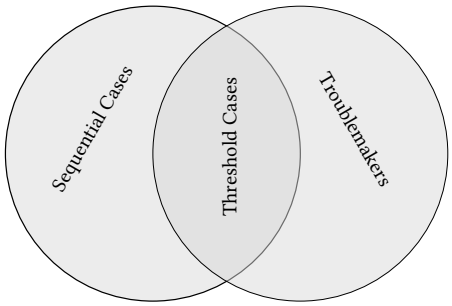


Figure 3.7: Threshold Cases are Sequential Cases that are Troublemakers.

Even though there is no single generally accepted definition, here is a rough-and-ready characterization of the two concepts that suffices for the context of this work. *Overdetermination* occurs when an effect has multiple sufficient causes, any of which would be enough on its own to bring about the effect, and they are all actual (i.e., they all occur). Thus, in overdetermination, multiple actual causes independently suffice to bring about an effect, and they all actually occur. Each cause is, in a sense, redundant; due to the other causes, the effect would have happened without it. *Preemption* occurs when a potential cause is rendered non-actual or irrelevant by another, prior or more direct, cause. In other words, while multiple factors might have been sufficient to bring about the effect, one specific factor effectively 'takes over' and becomes the actual cause, thereby preempting the others. Thus, in preemption, one actual cause effectively nullifies another potential cause, making the latter irrelevant in the actual bringing about of the effect.

This manifold overdetermination can be illustrated particularly vividly using Chicken Counter as an example. Given that 81 chicken purchases have been made, the breeding and treatment of chickens are indeed overdetermined by the actions of all the customers collectively. The causal chain that leads from chicken purchases to chicken breeding is activated many times over by the collective actions of these customers. Hence, the overall process of chicken ordering is not sensitive to the actions of any one cus-

[58] Shelly Kagan calls them "trigger cases" (cf. Kagan 2011).

tomer. This indeed creates a situation where individual inefficacy arises from overdetermination—each individual's action is rendered inefficacious by the sufficiency of other individuals' actions for the same result.

George's situation in Job Market, however, does not perfectly fit the typical mold of preemption (but the modified version does). George's potential refusal of the job wouldn't prevent the job from being done because Paul would take it in his stead. But George taking the job *and* Paul taking the job wouldn't have the same consequences, i.e., both would have different effects. However, we might stretch the term to capture the underlying structure in Job Market. George's refusal is rendered causally inert by Paul's willingness to take the job. After all, his willingness functions as a (genuinely potential) preemptive cause for the continuation of the research, even in a possibly more dangerous direction, regardless of George's decision. This effectively 'bypasses' the potential for the best consequence that might have been actualized if none of them took the job. George's potential to bring about the best consequence is nullified by Paul's readiness to step in, in much the same way that a (classical) preemptive cause nullifies the effect of another potential cause. This highlights an interesting aspect of the case: George's moral stand has a kind of 'fragility' akin to the fragility of preempted causes—it doesn't seem to make a difference due to the structure of the situation, much like a preempted cause doesn't get to make a difference due to the structure of its causal network. This brings out the tragic element of George's situation, emphasizing the difficulty of making a meaningful moral difference in a world where others stand ready to nullify our good intentions.

I thus believe that the distinction between overdetermination and preemption illuminates the structural differences between cases like Chicken Counter and Job Market. Beyond this conceptual acuity, however, the distinction does not play a vital role in my project, as it attempts to solve both types of structures using the same approach. Following others (Jackson 1987, 1997; Spiekermann 2014; Zamir 2001), will therefore use the term Overdetermination Cases as a generic label for both types of cases.

Compare that to our running example Two Factories. Recall:

Case 3.1 (Two Factories) *Ann and Ben each own a factory along the same river. Both can choose to produce either cleanly or cheaply, thereby polluting the river. The local market is highly competitive. Thus, a factory that produces cleanly would become noncompetitive if the other factory pollutes. The local social system is underdeveloped, and the economic situation is terrible. Hence, if a factory closes, this will cause significant unemployment and social hardship. If at least one factory produces cheaply, the resulting pollution will eventually destroy the local ecosystem and erode the livelihood of a village downstream. However, any additional polluter would not make the situation worse in this regard. Ann and Ben decide and act independently in the sense that neither of*

them can coordinate with the other, nor can any agent observe the actions of the other. Thus, whatever each agent does, they would do it regardless of what the other does.

If we (plausibly) assume that if both pollute, at least one of the two does something intuitively wrong (and for symmetry reasons, we should then assume that both are acting wrongly because the situations of Ann and Ben do not differ in any relevant aspects), Two Factories is a No-Difference Case.

However, in contrast to Job Market and Chicken Counter, Two Factories leaves the order in which the agents' actions occur unspecified. I call cases like Two Factories, where the order of action is not specified, Coordination Cases. Two Factories cannot be represented in extensive form.[59] Instead, we can represent it in what is called a *normal form*, i.e., in tabular form:

		Ben	
		pollute	produce cleanly
Ann	pollute	second-worst	worst
	produce cleanly	worst	best

As we know, Two Factories is a Troublemaker—and so it has a structure that is quite similar to Threshold Cases: Depending on the temporal perspective we take, Ann and Ben can truthfully utter different variations of (6) assuming that both will pollute, are polluting, or have already polluted. In this respect, Job Market, Kagan's Chicken Counter, and Two Factories all apparently fall into the same category according to Glover's classification—they are all Overdetermination Cases.

However, the difference to Threshold Cases lies in how the two agents in Coordination Cases appear to mutually overdetermine each other's actions. To quote Michael J. Zimmerman again (1996, p. 257), it seems as if in such cases there "is a sense in which [...] two wrongs [...] make a right". I will use the term Mutual Exculpation Cases to refer to Coordination Cases that are also Troublemakers. Figure 3.8 shows this rela-

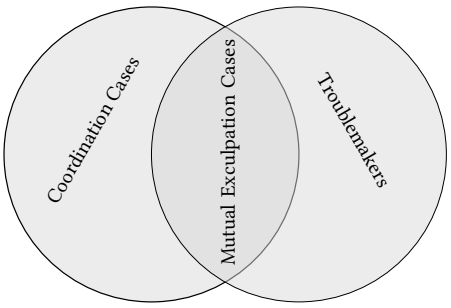


Figure 3.8: Mutual Exculpation Cases are Coordination Cases that are Troublemakers.

[59] If one leaves the rather exotic borderline case of actual, simultaneous actions aside, one can represent Two Factories as a set of two 'possible' extensive forms, for example. Felix Pinkert (2015, p. 975, Fig. 2) has suggested this.

tionship. In contrast, Figure 3.9 gives an overview over the 'space' of collective decision situations, without involving the Troublemaker question.

Two remarks on Coordination Cases. First, like Threshold Cases, these cases can involve an arbitrary number of agents. Here is an example with a slightly xenophobic connotation from R. B. Brandt (1959, p. 389) that pitches the Challenge (arguably as No-Difference Challenge) to make a case for Brandt's rule-utilitarianism and that still sounds pretty contemporary:[60]

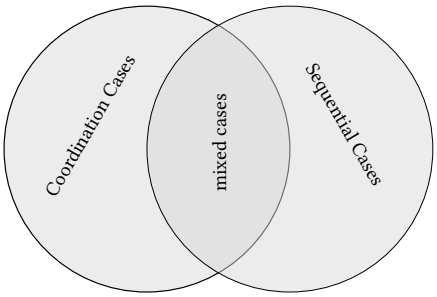


Figure 3.9: Collective decision situations can be Coordination Cases or Sequential Cases, depending on whether the order of actions is fixed or open. Mixed cases, where the order is fixed for some but not all actions, are possible. I omit them in this project as they are normally ignored in the literature, but that should not worry us regarding generality. Whatever we learn with respect to 'pure' cases should be easily transferable to mixed cases.

> Suppose that, in wartime England, people are requested, as a measure essential for the war effort, to conserve electricity and gas by having a maximum temperature of 50 degrees F. in their homes. A utilitarian Frenchman living in England at the time, however, argues as follows: "All the good moral British obviously will pay scrupulous attention to conforming with this request. The war effort is sure not to suffer from a shortage of electricity and gas. Now, it will make no difference to the war effort whether I personally use a bit more gas, but it will make a great deal of difference to my comfort. So, since the public welfare will be maximized by my using gas to keep the temperature up to 70 degrees F. in my home, it is my duty to use the gas."
>
> According to the act-utilitarian theory, this argument is perfectly valid. But we should not take it seriously in fact. Why not? At least part of the reason is that we think that, if a sacrifice has to be made for the public good, all should share in it equally. Imagine the outcry in Britain, if it became known that members of the Cabinet, who knew that electricity and gas were in good supply because of the country's willingness to sacrifice, used this argument to justify using whatever power was necessary to keep their homes comfortable.

In this case, it is irrelevant and, above all, unspecified who sets their heating to which level and in what order. It is, thus, a Coordination Case.

[60] Not only because it restates once again a variant of the Challenge (I'd say in the form of the No-Difference Challenge), but also because, living at the time of another actual war in Europe, I at least have heard that argument more than once in 2022, when the German Gas reserves where considered insufficient for the winter (cf. specifically Habeck's statement and the surrounding debate mentioned in the Preface, page v).

Second, although Coordination Cases clearly share structural similarities with prisoner's dilemmas, they are *not* to be confused with what is sometimes called "moral prisoner's dilemmas" (Parfit 1984, chapter 2). While these, like Coordination Cases, leave the order of actions unspecified, they presuppose agent-relative obligations or values (or, in Parfit's terms, "aims"), which, in fact, gives rise to the typical structure of a prisoner's dilemma, with 'moral payoffs' for each agent.

Besides being an example for a *n*-person Coordination Case, Brandt's Frenchman example is also a case where (7) can apparently be uttered truthfully by the involved agents and thus brings us back to Glover's distinction. Such cases involve actions that, in some sense, make no difference at all or at most a negligible one. In sum, however, many such acts do matter a lot.

The most important real-life candidate for such a case has already been mentioned in the introduction: the man-made climate crisis. The individual contribution of a normal citizen aggregated over a whole lifetime is probably measurable and statistically quantifiable. Still, this at most concerns what Julia Nefsky (2011) called the "underlying dimension" (cf. Kagan 2011): whether one single life is lived in an ascetic or wasteful manner probably makes no *moral* difference with regard to climate, even if it might be measurable in some way. Such real-life cases, however, bring into play many empirical questions that we should leave either to climate scientists or even cognitive psychologists (cf. E. N. Dzhafarov and D. D. Dzhafarov 2010b).

Thankfully, Glover provides an example of his own that avoids reliance on empirical questions. While Glover formulates his example in the form of one long description of two successive resolutions of two structurally identical collective decision-making situations, it proves expedient to pull these two sub-cases explicitly apart. This way, we get what Derek Parfit once called a "Glover pair" in an unpublished manuscript (Parfit 1988).[61] First, consider (Glover and Scott-Taggart 1975, pp. 174–175)

Case 3.4 (Beans and Bandits– One to One) *Imagine a village with 100 very hungry, nearly starving tribesmen, each preparing their lunch on one of 100 small fireplaces. 100 mildly hungry bandits are waiting outside the village for the right moment to steal the villagers' food. While the villagers are briefly distracted and turn their backs on their fireplaces, the bandits sneak into the village unnoticed. Each thief steals one villager's bowl to satisfy their appetite.*

Certainly, *no* bandit can truthfully utter (7). Each and every bandit's action makes a difference for the worse since it costs a villager their well-deserved and much-needed lunch. Hence not stealing a bowl would, no doubt, make a difference for the better. But now compare this to the second case:

[61] I also simplified the description a bit.

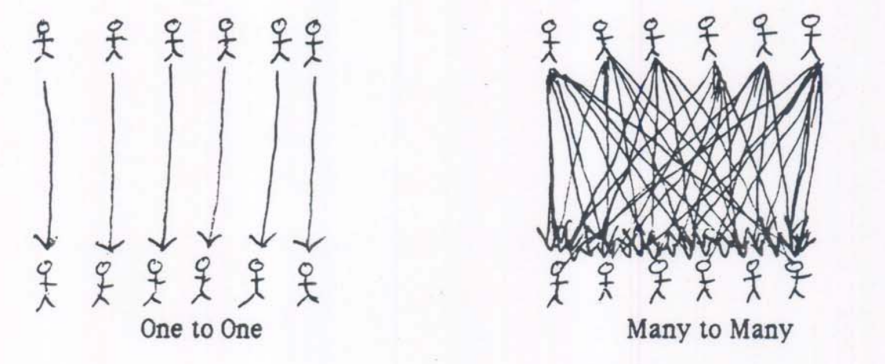


Figure 3.10: Original drawings by Derek Parfit (1988, p. 29) illustrating the structure of both Beans and Bandits situations. The bandits are depicted by the stick figures at the top, the villagers by those at the bottom, and arrows indicate acts of stealing (where the sum of the effects of all actions directed at one villager in Many to Many equals the effects of one action directed at that villager in One to One).

Case 3.5 (Beans and Bandits– Many to Many) *Imagine a village with 100 very hungry, nearly starving tribesmen, each preparing their lunch on one of 100 small fireplaces. 100 mildly hungry bandits are waiting outside the village for the right moment to steal the villagers' food. While the villagers are briefly distracted and turn their backs on their fireplaces, the bandits sneak into the village unnoticed. Each thief steals one bean from each villager's bowl to satisfy their appetite.*

Assuming that one bean more or less makes no moral difference to the villagers—no matter how many beans are left—every single one of the 100 bean thefts committed by each bandit is 'morally inefficacious' in the relevant sense. Then, when uttering (7), each bandit would say a truth. Thus, the second version of Beans and Bandits is another candidate for a No-Difference Case, at least if we assume that stealing the beans is intuitively morally wrong. Thanks to its collective context and general structure, this Beans and Bandits case is also a Troublemaker. Figure 3.10 shows Parfit's illustration of both situations' structure (cf. Parfit 1988).

Other well-known examples of this class of Troublemakers have been introduced by Derek Parfit, and especially his Harmless Torturers and Drops of Water example still enjoy some popularity (cf. Parfit 1984). Let us have a look at a modified variant of Kagan's revised version of Harmless Torturers (cf. Kagan 2011, p. 116; see also Figure 3.11):

Case 3.6 (Harmless Torturers) *Carl is wired to a torture machine with a thousand identical switches. When none of the switches are flipped, no current runs through the machine, so Carl is in no pain. If all thousand switches are flipped, then a considerable current runs through the machine, and Carl is in tremendous pain (but no permanent damage is done to his body). But the flipping*

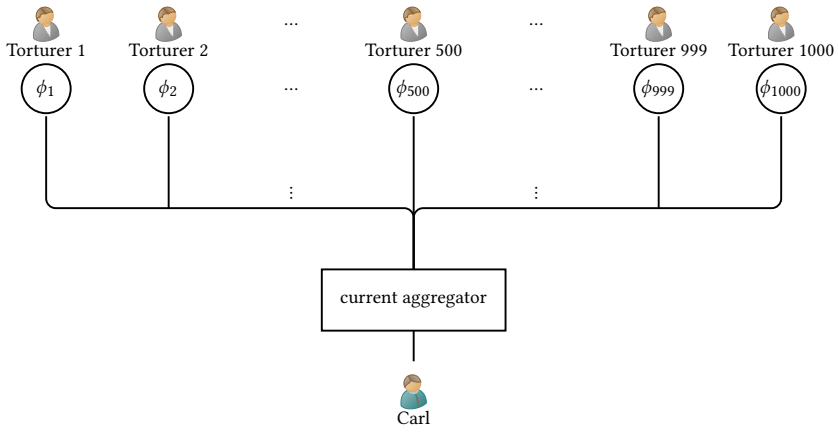


Figure 3.11: The setup of the Harmless Torturers case: 1000 torturers can each flip their switches ϕ_1 to ϕ_{1000}; the number of switches flipped determines the strength of the shock Carl receives.

of any given switch increases the current only by a tiny amount (well below the perceptually discriminable threshold for pain) so that Carl simply cannot tell whether one switch more or less has been flipped—regardless of how many other switches have already been flipped. Finally, imagine that a thousand different people each control a single switch and must decide whether to flip it or not. None of them cares about Carl or feels any remorse, but each of them enjoys flipping switches.

Of course, Harmless Torturers is a No-Difference Case (and also a Troublemaker). For if a sufficient number of torturers flip their switches, the resulting pain Carl experiences will outweigh the sum of the slight pleasures of the torturers in flipping their switches. It is evident that in this class of cases, it is relevant that, in some way, a set of individually negligible effects at an underlying dimension (cf. Nefsky 2011) somehow add up to a total effect that outweighs the sum of all benefits arising from the individual actions. Therefore, this class of Troublemakers cases shall be called Cumulative Effects Cases in the following. (Pure) Cumulative Effects Cases are *no* Overdetermination Cases. Figure 3.12 shows the relationship between the different cases.

We can now summarize what we have learned about different kinds of Troublemakers. First, Overdetermination Cases are those cases in which Individual Optimality is due to some form of overdetermination (or preemption), which is indicated by the agents' ability to truthfully utter (6). Second, Threshold Cases are Overdetermination Cases in which the order of actions is fixed, i.e., Sequential Cases that are Troublemakers. Third, Mutual Exculpation Cases are Overdetermination Cases in which the order of actions is *not* predetermined, i.e., Coordination Cases that are Troublemakers. Of course, mixed forms are conceivable, i.e., situations in which the order is fixed for

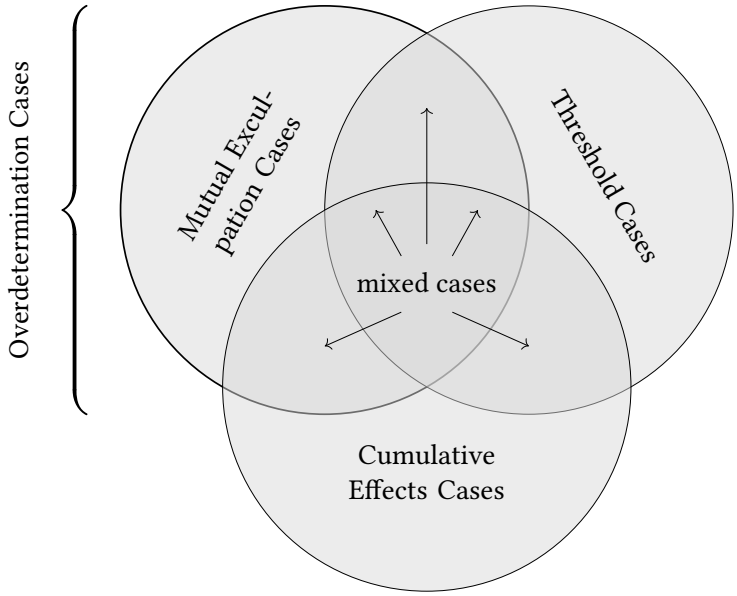


Figure 3.12: The set of all Troublemakers as the union of the three different types of Troublemakers. Overdetermination Cases are the union of Mutual Exculpation Cases and Threshold Cases. All mixed cases, i.e., all cases in any intersections, are omitted for systematic reasons, both in the literature and in this project. The idea is that, on the one hand, one can master the Challenge more easily for 'pure Troublemakers' and, on the other hand, once one has working approaches for all these pure type cases, one can also construct complex solutions for mixed cases.

some actions but not for all. Finally, there are Cumulative Effects Cases and, again, mixed cases with all other kinds of Troublemakers.

The rough-and-ready taxonomy of Troublemakers presented here proves useful in my project, primarily because the underlying different Individual Optimality true-making structures by which these types are characterized arguably call for different approaches to the Challenge. This divide-and-conquer strategy goes back at least to Derek Parfit (1984, chapter 3), but continues to enjoy great popularity today (see, for example, Kagan 2011). However, my taxonomy is, in its entirety (cf. Figure 3.3, Figure 3.7, and Figure 3.8), more nuanced.

For now, however, we can leave behind the taxonomy and return again to the No-Difference Challenge in order to wrap up this section. In a nutshell, the idea of the Challenge as a No-Difference Challenge can be summarized as follows: there are collective decision situations where some action is, intuitively speaking, morally wrong. However, because of individual inefficacy, i.e., the inability of the individuals to make a difference for the better by unilaterally acting otherwise, plausible difference-making

views apparently fail to come to the intuitively correct assessment.[62] This version of the CHALLENGE as a NO-DIFFERENCE CHALLENGE arguably finds an appropriate representation in the following argument:

The NO-DIFFERENCE CHALLENGE ARGUMENT – tentative

$P_{\exists \text{ NDCs}}$: There are No-Difference Cases: collective decision situations in which there is at least one agent who can act in a way such that it seems intuitively morally wrong, but the agent could not make a difference for the morally better by unilaterally acting differently.

P_{MOCOR}: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

$P_{\neg\text{intu}}$: If a moral theory assesses intuitively morally wrong actions as morally right for a significant class of situations, then T is counterintuitive.

$C_{\neg\text{intu}}$: MOAC is counterintuitive.

This argument is apparently deductively valid,[63] so we directly turn to the premises. In light of the example cases presented, $P_{\exists \text{ NDCs}}$ seems reasonably well-supported, at least if we accept, for the sake of argument, the intuitive assessments involved. Thus, the soundness of the argument hinges on whether P_{MOCOR} and $P_{\neg\text{intu}}$ are true. Given MOCOR, it seems indeed hard to deny that P_{MOCOR} is true, but we investigate this in detail in the next subsection (and, to be honest, in the second part in even more detail) when introducing the INTERNAL CHALLENGE and, thus, the ARGUMENT. $P_{\neg\text{intu}}$ seems a lot like

[62] Note that I have restricted the investigation here to the *negative* formulation of the NO-DIFFERENCE CHALLENGE. Alternatively, we could formulate it positively. There are apparently cases where some action does not make a difference for the *better*, and yet we seem to have a strong intuition that this action is *right*. This positive formulation, to my knowledge, is seldom discussed in academic circles. However, it often surfaces in everyday moral discourse. Consider, for instance, individuals who adopt veganism or consciously reduce behaviors tied to high carbon dioxide emissions. They believe these choices are morally right, even if they might not see an immediate tangible impact. One could reframe the NO-DIFFERENCE CHALLENGE to highlight the potential misjudgment of these intuitive stances. Indeed, academic discussions might exhibit a bias, emphasizing the risks of false positives (mistakenly deeming intuitively wrong actions as right) over the potential pitfalls of false negatives (overlooking actions that are intuitively right but are assessed as wrong).

[63] Assuming that the class of No-Difference Cases qualifies as a significant class of situations, which seems more than plausible.

a conceptual truth, even though one may question whether *one* clash with intuitions should suffice for labeling a whole theory counterintuitive. However, since the No-Difference Challenge makes a rather systematic point and diagnoses such a clash for a broad class of cases, we might accept it nevertheless. In this respect, the No-Difference Challenge Argument seems relatively convincing—as long as one shares the central intuitions, anyway.

At the same time, its appeal to intuitive assessments is also the weakness of the No-Difference Challenge. Since it attacks the intuitive extensional adequacy of a theory, it is, on the one hand, quite broad in its application. It threatens the convictions of any adherent of a theory T that (in possibly slightly modified No-Difference Cases) comes to the relevant assessments, as long as he shares the central intuitions. As we shall see, it can, therefore, also be brought to bear against subjective varieties of act-consequentialism. On the other hand, one can escape the Challenge as No-Difference Challenge if one simply does not share the central intuitions. And even if one does share them, one might be willing to bite the bullet: maybe we are learning here, drawing an important general lesson for our endeavor of theory-building. What theory comes without any intuitively questionable judgments?

In other words, if we are to formulate a stronger version of the Challenge that puts moac in more serious trouble, we could do so as soon as we provide reasons why these seemingly intuitive assessments cannot simply be written off by moac's advocates. Suppose it can be shown that moac is even *committed* to these assessments. In that case, we have an existential challenge because then the diagnosed inadequacy of moac would not be that of an intuitive extensional inadequacy but that of a theoretical inconsistency—which is clearly an unbearable burden for any theory. As outlined in the introduction, there is indeed such a theory-internal framing of the Challenge.

3.5 The Challenge as Internal Challenge

A much stronger formulation of the Challenge describes it as an intra-theoretical conflict. I have already sketched this formulation in the introduction – in this section, I will dig a little deeper.

3.5.1 Starting from a Non-Starter: The Principle of False Universalization

One of the earliest modern analytic engagements with Troublemaker-like cases is found in a 1916 essay by C. D. Broad (1916). Like Glover, Broad grap-

ples with a particular structure of argumentation that often appears in this context, viz.,

(8) What if everyone did that?

A brief examination of Broad's considerations proves useful for this project in two ways. First, (8) stands for a practice of justification often put forward in the context of collective action. As we will see, Broad argues convincingly that this practice is unsuitable from a consequentialist perspective; therefore, we can also set it aside. Second, Broad's discussion of this justification practice introduces in passing a version of the principle MH and is thus ideally suited for the transition to a stronger framing of the CHALLENGE as INTERNAL CHALLENGE, which heavily relies on PMH. Broad made these points almost 50 years before Glover prompted the debate surrounding the NO-DIFFERENCE CHALLENGE. It is worthwhile to turn briefly to Broad's reflections.

Considerations akin to (8) typically end with the condemnation of the action under consideration because, if everyone were doing the action under consideration in such circumstances, the consequences would be unacceptable. C. D. Broad (1916, p. 377) pondered on that practice by discussing the specific principle instantiated in such considerations:

> A man proposes to himself a certain course of action and debates whether it be right or wrong. At a certain stage he will say to himself, or, if he be discussing the matter with a friend, his friend will say: Suppose *everybody* did what you propose to do. The consequences of this hypothesis will then be considered, and, if they be found to be bad, the man will generally consider that this fact tends to prove that his proposed action is wrong.

Broad called that principle the *principle of false universalization*. It is a principle of *false* universalization, as we "are asked to believe that the rightness or wrongness of many of our actions depends on the probable consequences, not of what we judge to be true, but of what we know to be false" (ibid.).[64] Given an *actual* individual decision situation, the principle invites us to consider a *hypothetical* collective decision situation where *all* the (hypothetical) agents perform the (hypothetical) action corresponding to the (actual) action the (actual) agent is considering. Then, the (hypothetical) outcome of everyone—or at least a sufficient number of people to produce some relevant

[64] Broad certainly is correct when diagnosing a certain similarity between the principle of false universalization and Kant's categorical imperative (especially in the so-called law of nature formulation). But there can be no doubt either that there are important differences. Thus all that follows is not to be read as an argument against Kant's theory. Also, note that there are many points of connection to universalizability and impartiality requirements involved in many schools and traditions of moral theorizing.

effect—performing 'the action' under consideration is evaluated. If the result is judged to be unacceptable (or "bad" in Broad's own terms), then the original action under consideration is deemed wrong. Here is what I take to be a fair formulation of that principle:

Principle 3.3 (False Universalization (negative)) *Let ϕ be an option that would by itself not make a difference for the worse. Option ϕ is morally wrong if a sufficient number of type-identical actions would together produce morally unacceptable consequences.*

Broad considers several examples and several possible ways of applying the principle. Here is one particularly interesting passage (ibid., pp. 383–384):

> I walk through a field and pluck an ear of corn. Is this right, wrong or indifferent? If I now say: Suppose a million people walked through and each plucked an ear, the results would be very bad [...].[It] seems to me that the argument from the damage done by a million ears being plucked to that done by the plucking of one is most precarious. The consequences that have to be considered cannot be the mere separation of the ears from the stalk; this, like all physical events, is in itself morally indifferent. We obviously have to go further and consider the effects on the state of mind of the owner of the field and of others. Now it seems perfectly possible that no one's state of mind is in the least better or worse for the plucking of one ear and yet that it may be very much the worse for the plucking of a million. There is absolutely no logical reason against this and it seems to me to be true. The most probable account of the matter is that the plucking of a certain finite number *n* (varying of course with the circumstances) is absolutely indifferent, while the plucking of any greater number leads to consequences which get worse as the number gets greater.

There is obviously a connection between Troublemakers and the cases Broad targets here, even though his cases are *not* Troublemakers. In Broad's cases, a single individual act makes no relevant difference, but a multiplicity of *hypothetical* acts would have unacceptable consequences. Next, the False Universalization invites to use these hypothetical consequences as basis for condemning the actual, individual option under consideration. In Troublemakers, by contrast, a multiplicity of *actual* acts taken together have unacceptable consequences, while the effects of the individual acts seem negligible *because*, given the other actual acts, none of them seems to make any significant difference.

What would help with Troublemakers, to arrive at the apparently desired assessments, is thus a principle rather like this one:

Principle 3.4 (Totum pro Parte) *Let ϕ be an option that would by itself not make a difference for the worse. Option ϕ is morally wrong if it is part of a combination of actions that produce morally unacceptable consequences.*

Totum pro Parte would allow us to assess the individual options in Troublemakers as wrong even though, by definition, no alternative would have made a difference for the better. Of course, this approach to the CHALLENGE has been considered, and we will briefly examine this strategy, as proposed by Frank Jackson (1987), along with its drawbacks in Chapter 4, before ultimately dismissing it.

False Universalization has its own problems. Most importantly and in contrast to Totum pro Parte, it links the moral status of an *actual* option to the *hypothetical* outcome of certain *hypothetical* combinations of actions. Especially from an objective consequentialist perspective, it is difficult to see how *merely hypothetical* consequences should carry 'moral weight'. As a result, the principle cannot (or at least not readily) be made fruitful for MOAC. Indeed, Broad agrees that the principle of false universalization fails exactly because it fails, if one may say so, to establish a robust link between assessed options and the potential 'wrongmakers'. More than that and adopting a rather consequentialist perspective, he even concludes "that both on practical and ethical grounds it is most unlikely that you can ever safely argue from the goodness or badness of the effect of a number of precisely similar acts to the rightness or wrongness of a single act of the class" (Broad 1916, p. 385). (Note that if this were true, not only False Universalization but also Totum pro Parte would be a non-starter.)

Broad's considerations, however, come with a caveat that brings us back to PMH and the collective reading of Congruence, namely the idea that right actions go hand in hand with moral optimal outcomes. Because in exactly this sense Broad discusses an *inverted* variant of False Universalization that is meant to be used to establish not the wrongness of an action but its *rightness*. In this variant, not the bad effects of a large number of hypothetical actions are used as a basis for an assessment of an individual option under consideration, but the *positive* effects of a multitude of hypothetical actions are taken into account (or, respectively, the negative effects of a continuation of the status quo) for such an assessment. In other words, according to that inverted principle, we infer the rightness of each such option from the morally welcome consequences that a multitude of certain *hypothetical* actions (of the same type) as the option under consideration would have. Consider

Principle 3.5 (False Universalization (positive)) *Let ϕ be an option that would not make a significant difference for the better. Option ϕ is morally right if a sufficient number of type-identical actions would produce morally acceptable (which, in the spirit of MOAC, may well mean: optimal) consequences.*

In the wild, we find the application of the positive variant at least as often as that of the negative, especially in the context of large-scale challenges such as climate change. Here the principle is often applied not only in the private

sphere but in the public discourse as well. Where the negative version is used to tell us to give up our beloved steaks, the positive version is used to reinforce our consumption of oat milk. The two variants, apart from their lack of theoretical persuasiveness (at least from a consequentialist perspective), complement each other excellently.[65]

For some reason, Broad believes this positive variant of the principle holds (ibid., p. 391):

> Is there then no valid use for the principle of false universalisation in ethics? I think there is at least one, though it is a very modest one. It can be used to refute a certain kind of mistaken judgment about the rightness of a suggested act. Suppose that certain acts are very unpleasant to everyone and entail very real sacrifices from which everyone shrinks. Suppose further that the performance of such acts by a certain number of persons is essential to the attainment of a considerable good or the avoidance of a considerable evil. If now a man says: I will not act thus *because* I dislike the sacrifice then it is open to us to point out to him that, if this be his sole ground, it is just as valid a ground for all other people, since by hypotheses they all dislike the sacrifice. If then he is right in refusing to do the act, all other people will also be right in refusing on the same ground. But the result will be that a great good will be lost or a great evil suffered.

It is very interesting in the context of this thesis to understand why Broad believes that he and the consequentialists *need* this principle. After all, it seems unfounded to accept the relevance of hypothetical combinations in one case but not the other. Thankfully, Broad lets us know why he *wants* the positive variant to be true (ibid., p. 392):

> Now *it cannot be the case that the result of a number of right actions can be a state of affairs which can be foreseen to be worse than if people had acted differently* [emphasis added]. Hence we can conclude that these actions could not all be right. But if his ground for supposing that his action was right were valid all these actions would be right.

This means that Broad proposes here to endorse the positive principle of false universalization because he thinks that otherwise another principle he believes to be immutably true would be violated: that it cannot be the case that some combination of exclusively right actions leads to predictably suboptimal outcomes. This, of course, is a variant of PMH from the introduction.

Before I explain that, let me emphasize that I do not think that Broad can have it both ways. He claimed (and I think rightly so) that the negative variant of False Universalization is invalid because we are not allowed to cite the

[65] There is certainly a connection to what I called the positive variant of the No-Difference Challenge, see footnote 62.

unacceptable *hypothetical* consequences of *hypothetical* actions in support of the condemnation of *actual* individual inefficacious options, since there is a missing link between the hypothetical and the actual. But for the very same reasons, then, we are not allowed to cite the foreseeable *hypothetical* betterness of consequences of *hypothetical* actions in support of the rightness of *actual* individual options that come with certain moral costs.

I am not here to defend or attack Broad's reasoning, though. For my project, the important point is this: Broad considers it beyond doubt that "it cannot be the case that the result of a number of right actions can be a state of affairs which can be foreseen to be worse than if people had acted differently." In the service of defending *this* idea he is willing to accept the principle of false universalization for certain cases. The idea can arguably be stated as

Property 3.1 (Broad's Property) *If all agents do right, then, necessarily, they do not foreseeably produce morally suboptimal consequences together.*

I call Broad's Property a *property* (and not, say, a principle) because it is best understood as the description of a property of moral theories. This property may (or may not) accrue to some moral theories and, of course, involves an implicit universal quantification. That a moral theory has the corresponding property means that, for all decision situations (within its domain) and under all relevant contexts, this theory makes recommendations such that following these recommendations, i.e., doing what is right according to the theory in these situations and relative to these contexts, can not lead to foreseeably suboptimal outcomes.

In other words, Broad believes that, if all agents act rightly, then they are guaranteed to either together produce a morally optimal outcome or they together *unforeseeably* produce a morally suboptimal one. Hence, Broad's formulation has an epistemic and, thus, subjective twist: The phrase "can be foreseen" immediately makes one wonder for *whom* it should be foreseeable. From Broad's example, it is clear that he has in mind the agents' perspectives. However, we might want to use Epistemic Limes to translate Broad's property into an objective stance. Recall:

Principle 2.4 (Epistemic Limes) *Let T_{o} be an adequate objective (i.e., non-epistemic) moral theory and let T_{s} be an adequate subjective (i.e., epistemic) moral theory. For a given decision situation involving an agent A with a set of options Φ and a set of relevant contexts C: If A knows all relevant facts and has no incorrect relevant beliefs, then*

$$T_{\mathrm{o}}, D, C \vDash \mathrm{R}\phi \quad \textit{if and only if} \quad T_{\mathrm{s}}, D, C \vDash \mathrm{R}\phi.$$

Although Epistemic Limes must be taken with a grain of salt—as there might be things one cannot know at the time of acting, such as what other

agents are doing or will do—we may, for a moment, assume that we can apply it to drop the subjective aspect, giving us an objectivist's version of Broad's Property. This version then boils down to the collective reading of Congruence from the introduction:

Principle 1.1 (Collectively Maximizing – tentative)
If all agents act rightly, then they are guaranteed to produce the morally best outcome they could together bring about.

Thus, we have indeed re-arrived at what I earlier called, following Feldman (1980), the Principle of Moral Harmony (PMH). When Broad writes that "it cannot be the case" that MH is violated, he implicitly says that he take MH as a necessary property of every plausible moral theory. In other words, Broad would have apparently accepted:

Criterion 1.2 (Moral Harmony (MH) – tentative) *A moral theory is adequate only if it is true that if all agents act rightly (according to this theory), then they are guaranteed to produce the morally best outcome they could bring about together.*

At this point, let me draw three lessons from Broad: first, consequentialists must not hope that False Universalization will help to master the Challenge; second, one of the fundamental principles underlying the Challenge is much older than the discussion of the Challenge itself; third, this principle is not unique to genuinely consequentialist moral thought (as Broad was no consequentialist; cf. Gustavsson 2021).

The following subsection will be devoted to the explicit reconstruction of the Argument and thus of the foundation of the Challenge as Internal Challenge, its strongest variant.

3.5.2 Regan's Impossibility Result

Donald Regan was not one afraid of bold claims. Here is what he sets forth as the goal of his book *Utilitarianism and Co-operation* (Regan 1980, p. vii): "In this essay I shall first analyze and then dissolve a contradiction which the existing literature suggests is inherent in utilitarian theory and which, if it were genuinely indissoluble, would weigh heavily against the acceptability of any form of utilitarianism." As my present project is based on the existence of that apparent contradiction, it's important both that it can be resolved and that it has not been resolved successfully by Regan in 1980. So we should follow Regan's thoughts in quite some level of detail. He sketches his plans further (ibid., pp. vii–viii):

> [T]here are two distinct and equally compelling particular intuitions subsumed under the general utilitarian intuition that moral agents should be

> required to maximize good consequences. According to one of these particular intuitions, each individual agent should be required to act in such a way that the consequences of his own behaviour are the best possible in the circumstances confronting him as an individual. According to the other of these particular intuitions, any group of agents should be required to act in such a way that the consequences of their collective behaviour are the best possible in the circumstances confronting the group as a whole. [...]
>
> In the course of this essay, I shall show both that the problem I have just indicated is a real problem, and that there is a solution. In the first half of the essay, I shall demonstrate that the two particular intuitions I have identified [...] can not be reconciled by any moral theory of the general sort which utilitarian theorists have proposed up to now.

It should be obvious that the two allegedly 'irreconcilable intuitions' correspond to MOCOR and Collectively Maximizing, while the "general utilitarian intuition" refers to what I called

View 1.1 (Congruence) *The right and the best are congruent in the sense that doing what is right goes (necessarily) hand in hand with bringing about the morally best consequences that can be brought about.*

Regan claims that he demonstrates that "the two particular intuitions [...] can not be reconciled". It is this 'impossibility result' that corresponds to the INTERNAL CHALLENGE as sketched in the introduction and that shall be reconstructed, partially in combination with similar arguments by other authors, in more detail in this section.

It should be noted that Regan does not target merely MOAC theories but a broader set of 'victims'. To understand the scope of his ambition, some of Regan's technical lingo proves helpful. Most importantly, he distinguishes "exclusive act-orientation" theories—corresponding to what he referred to as "the general sort" in the previously cited quote—from all other approaches to moral theorizing. While Regan sees himself unable "to produce a definition of exclusive act-orientation which is both precise and completely general" (Regan 1980, p. 109), he gives a characterization of the property for the context of Coordination Cases (ibid., p. 113):

> [W]hether an agent satisfies a traditional consequentialist theory depends, ordinarily, on what he does from a list of acts [...] and not on what he tries to do or how he decides what to do. This is the feature of traditional theories I refer to by saying they are "exclusively act-oriented".

Much more recently, Douglas Portmore offered a more general definition: "a theory is exclusively act-orientated if and only if it requires only that agents perform and refrain from performing certain voluntary acts" (Portmore 2018, p. 14). Whatever the best way to put the distinction, it should be clear that

MOAC indeed is such a theory by the very structure of MOCOR alone. But even if we can somehow save MOAC from Regan's attack, it is important to remember that it may similarly affect every other exclusively act-oriented theory. However, it will turn out that it is not at all trivial to get to the heart of the intuition behind this property.

While Regan's work did not attract a particularly wide audience, Derek Parfit made it the basis of one very influential chapter of his *opus magnum Reasons and Persons* (Parfit 1984).[66] Accordingly, Regan's work echoes via Parfit's and informs current works on the Challenge like Pinkert 2015; Portmore 2018, among others. To my knowledge, Regan's discussion of the Challenge is the most thorough up to this day and thus deserves a particularly close look in the context of this thesis. After all, as mentioned earlier, my project can be seen in some respects as a reissue of Regan's project.[67] Yet, as similar as our goals are—to analyze the Challenge and, more specifically, the Internal Challenge in detail and then to solve it—our results are different.

However, Regan's thoroughness comes at a price. Although Regan limits his investigation to only one subvariant of Troublemaker, namely Coordination Cases, his work draws heavily on self-defined notions, abbreviations and various case distinctions. For at least some of the distinctions, a rough and ready understanding is necessary in order to understand his main argument. Because we can learn much about the strong strand of the Challenge by examining Regan's reasoning carefully, I take some time to unfold his argument the preliminary version of which we have seen in the introduction:

The Argument – tentative

$P_{\exists\text{TROUBLE}}$: There are Troublemakers: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

[66] Another distinct yet related strand of literature where Regan's work resonates lies within a specific area of game theory. Michael Bacharach's work on the so-called "Hi-Lo Cases" (Bacharach 1999, 2006) is notably influenced by Regan. (Both Regan and Bacharach were significantly inspired by Thomas Schelling's analysis of conflicts and coordination problems (Schelling 1980; see also Regan 1980, pp. 133, 191, 198, 260, 265).) Thus, Regan's influence continues to be evident in contemporary discussions within the branch of game theory concerned with cooperation and 'team reasoning', as demonstrated in recent works, e.g., Gold and Colman 2020; Petersson 2017.

[67] Regan's project is also heavily inspired by Allan F. Gibbard's 1965 article, which has already been mentioned.

P_{MOCOR}: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

P_{MH}: If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (according to this theory).

$C_{\neg\text{ADEQ}}$: MOAC is not an adequate moral theory.

In what follows, I reconstruct this argument step by step, always starting with Regan but also looking beyond Regan as a primary source. This will be done in three steps, each corresponding to one of the premises (but in a different order than in the argument, reflecting Regan's own story-telling).

3.5.2.1 Step 1: Whiff and Poof and $P_{\exists\text{TROUBLE}}$

Regan puts a very simple and artificial example[68] at the center of his book (Regan 1980, p. 18):

Case 3.7 (Whiff and Poof) *Suppose that there are only two agents in the moral universe, called Whiff and Poof. Each has a button in front of him which he can push or not. If both Whiff and Poof push their buttons, the consequences will be such that the overall state of the world has a value of ten units. If neither Whiff nor Poof pushes his button, the consequences will be such that the overall state of the world has a value of 6 units. Finally, if one and only one of the pair pushes his button (and it does not matter who pushes and who does not), the consequences will be such that the overall state of the world has a value of 0 (zero) units. Neither agent, we assume, is in a position to influence the other's choice.*

Whiff and Poof is a Coordination Case that can be represented in the following normal form:

[68] As Regan attributes correctly, this example is basically identical to Alan F. Gibbard: "The situation will be as follows. Jones and Smith sit in their isolation booths with red push-buttons. If at 10:00 a.m. both are holding down their push-buttons, they receive cake and ice cream, which is intrinsically good. If only one of them is holding his push-button down, however, they both receive electric shocks, which is intrinsically bad. If neither of them is holding his button down, nothing happens", Gibbard 1965, p. 215; he uses it to make roughly the same point Regan makes, but with the goal of shedding light on the relation between rightness in act- and in rule-consequentialist terms.

		Poof	
		not-push	push
Whiff	not-push	6	0
	push	0	10

It should be obvious that Whiff and Poof is structurally equivalent to Two Factories in the sense that they have the same outcome profile. Consequently, Whiff and Poof is a Mutual Exculpation Case. For now, a merely intuitive understanding of the notion of identical outcome profiles will suffice: There is a bijective[69] mapping from combinations of actions in Whiff and Poof to combinations of actions Two Factories that conserves the ordering of the values of the corresponding outcomes. Ann and Ben both producing cleanly corresponds to Whiff and Poof both pushing, etc.[70] Thus, whatever we learn from Regan's argument for Whiff and Poof with respect to MOAC can be straightforwardly transferred to Two Factories – and, more generally, to Coordination Cases. After all, according to MOCOR, the order of the values of outcomes of actions is all that matters for their moral assessment (as we have defined in Definition 2.3 at the end of the preliminary chapter).

Both Whiff and Poof and Two Factories seem to offer coherent, consistent and sufficiently complete descriptions of collective decision situations. Thus, given the lightweight criterion for the existence of cases in the relevant sense (cf. Section 2.1, page 18) they apparently warrant the first premise of the Argument, viz.

($P_{\exists\text{TROUBLE}}$) There are Troublemakers: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

The next component of Regan's argument is considerably more interesting.

3.5.2.2 Step 2: PropCOP and P_{MH}

The second theoretical piece in Regan's (1980, pp. 4–5) puzzle is

Property 3.2 (PropCOP (Regan)) *If all agents satisfy T in all choice situations, then the class of all agents produce by their acts taken together the best consequences that they can possibly produce by any pattern of behavior.*

[69] As introduced in footnote 36 on page 35.

[70] It is not worthwhile to spell out this definition formally here. Later, we will have to do so for the sake of an important argument.

According to Regan, PropCOP[71] PropCOP is one of the "two particular intuitions [that] can not be reconciled" (see above). For Regan, PropCOP is a (at least historically) widely accepted criterion of adequacy for consequentialist theories.

We already stumbled upon a very similar idea when considering Broad's passage above, and one will stumble upon similar ideas again and again in centuries of literature on moral philosophy, especially in the context of consequentialist thinkers. The following is a short excursion into the corresponding history of ideas in order to ensure that the notion I offered above does justice to the general idea.

We start with Fred Feldman, who, in the very same year Regan published his book, published an article (Feldman 1980) on what he called

Principle 3.7 (Principle of Moral Harmony– Feldman) *When all the members of a social group do what they morally ought to do, the group as a whole does benefit more than it would have from the performance of any worse alternative set of actions.*

The similarity between Regan's PropCOP and Feldman's Principle of Moral Harmony is certainly striking. While Fred Feldman does not refer to Regan, he does mention a variety of other proponents of very similar expectations either of morality in general or specifically of consequentialist theories. For instance,[72] take Berkeley's view in his *Passive Obedience* (Berkeley 1929, p. 239). Operating under the assumption that judgments of rightness are grounded in a set of moral rules (or "precepts" in Berkeley's terms) that he called 'the law of nature' which express God's judgments, Berkeley claims that "the law of nature is a system of such rules or precepts as that, if they be all of them, at all times, in all places, and by all men observed, they will necessarily promote the well-being of mankind, so far as it is attainable by human actions." Furthermore, we already encountered Jeremy Bentham's take on the issue in the introduction. He took his *Principle of Utility*, "which approves or disapproves of every action whatsoever, according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question", to be "capable of being consistently pursued" and

[71] "Prop" stands for "Property" and "COP" stands for a certain theory that is meant to have the 'cooperative' feature articulated in PropCOP (Regan 1980, p. 85):

Principle 3.6 (COP) *An act is right if and only if it is prescribed (for the agent whose act is in question) by that universal prescription for action, the universal satisfaction of which would produce the best possible consequences.*

[72] The following list is heavily inspired by Feldman's 1980 discussion of the principle in said article.

believes that "it is but tautology to say that the more consistently it is pursued, the better it must ever be for humankind" (cf. Bentham 2007). In more modern times, we can find a similar idea in the work of Stephen Toulmin (1953, p. 137), who once wrote that "we can provisionally define [the 'function' of ethics] as being 'to correlate our feelings and behaviour in such a way as to make the fulfillment of everyone's aims and desires as far as possible compatible'." Feldman collected several other examples of philosophers that embraced *some* version of PMH, some of which I have not yet mentioned, most importantly Kurt Baier (1958), Hector-Neri Castañeda (1974), and J. L. Mackie (1977).

Even though Feldman himself tried to argue against this principle, the idea remains alive. For instance, when Michael J. Zimmerman discusses the concept of moral obligation, he writes[73] in his chapter on cooperation (cf. 1996, pp. 257–258)):

> Thus, given [an account of moral obligation that demands everyone to do the best one can, call it *T*], *neither* [Whiff] *nor* [Poof] does wrong in [not pushing], even though [...] the best that *both* can do is *not* done. [...] This implication of [*T*] should surely trouble anybody inclined to the view that one ought to do the best one can, inasmuch as it demonstrates that there is a sense in which universal satisfaction of [*T*] is compatible with the best that can be done *not* being done.

Similarly, Douglas Portmore (2018, p. 13) formulated the idea as an explicit criterion for the correctness of moral theories, drawing explicitly from Regan's ADAPTABILITY (to be discussed in the next section):

Criterion 3.1 (MH– Portmore) *A moral theory T is correct if and only if the agents who satisfy T, whoever and however numerous they may be, are guaranteed to produce the morally best world that they could together bring about.*

Finally, we can use the opportunity to return to Felix Pinkert—the inventor of Two Factories—who writes (Pinkert 2015, p. 975):[74]

> Act Consequentialism judges that both Ann and Ben act rightly by polluting, even though they together could easily have brought about much better outcomes by both producing cleanly. So if Consequentialists only have Act Consequentialism to morally apprise [sic!] a situation, then they let Ann and Ben off the hook for together producing collectively suboptimal outcomes. This is at odds with the following claim:

[73] Adapted in the following in order to match Whiff and Poof instead of his specific case.

[74] Note that I leave Pinkert's principle untouched in both naming and wording, but pull it out of the citation so that it can be numbered consecutively with the other definitions and principles.

Principle 3.8 (On-the-hook) *In any collection of agents who together gratuitously fail to bring about collectively optimal outcomes, there must be some relevant morally objectionable facts about some of the agents.*

Pinkert has particularly strong opinions with respect to On-the-hook (Pinkert 2015, pp. 976–977):

> [...] On-the-hook is a widely shared assumption in the philosophical discussion of Consequentialism and no-difference cases. [...] On-the-hook should not be understood as a specifically Consequentialist position. Instead, it should be understood as the contraposition of a second-order claim about morality in general and, hence, as a desideratum for any moral principle. According to this claim, the relation between morality and overall value is such that if everyone always satisfied all requirements posed on them by morality, the world would be as good as it can be (as far as agents' influence is concerned). [...]
>
> Thus understood, On-the-hook has considerable intuitive appeal, and a moral principle that can accommodate this intuition is, other things equal, strongly preferable to a moral principle that cannot accommodate it.

This second-order claim with "intuitive appeal" Pinkert evokes here is certainly *a* collective reading of Congruence, i.e., a version of PMH. In contrast to MH, Pinkert's On-the-hook allows compromises concerning the possible fulfillment of this second-order claim by adding the formulation "gratuitously". Pinkert writes (ibid., p. 976):

> By calling a failure to bring about optimal outcomes "gratuitous," I mean that the failure cannot be explained by mitigating circumstances due to which we could not expect a given group to collectively act optimally. Typically, such circumstances consist in non-culpable misinformation or lack of information. Since Ann and Ben know all relevant facts, I assume that their failure to bring about optimal outcomes is gratuitous.

Arguably, Pinkert is assuming here a principle in the spirit of Epistemic Limes (cf. Principle 2.4 on page 31), claiming that this would allow subjective accounts to embrace On-the-hook, too[75]—something that is implausible for MH to assume.

It is worth briefly exploring the qualification Pinkert has in mind and discussing why an unqualified formulation of MH is unappealing for subjective accounts (as I will argue later in this thesis that even objective theories ultimately require a qualified formulation of MH). For this purpose, we briefly leave the arena of MOAC theories and enter that of Maximizing Subjective Act-Consequentialism (MSAC).

[75] Pinkert's On-the-hook is thus a direct descendant of Broad's Property, cf. Subsection 3.5.1.

MSAC is a family of subjective consequentialist moral theories, i.e., all of its members come with a subjective stance (as introduced in Section 2.3, View 2.2) that takes into account the epistemic situation of agents.

As becomes clear by a glimpse at their criterion of rightness, the Maximizing Subjective Criterion of Rightness (short: MSCOR), they are *prospective* theories (as they base the assessment of actions on their expected future consequences):

Principle 3.9 (MSCOR (prototypical)) *It is right to perform a certain action if and only if there is no alternative with expectedly better consequences.*

Prospective theories like MSAC theories do certainly not *generally* embrace MH. It is helpful to have an example at hand that shows why this is the case. Consider the following, provided by Frank Jackson (1991, pp. 462–463):[76]

Case 3.8 (The Drug) *Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the literature has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it. One of the other two drugs, either B or C, will completely cure the skin condition; the other though will kill John, and there is no way that she can tell which of the two is the perfect cure and which is the killer drug.*

We can depict this case in the extensive form shown in Figure 3.13.

Even without concrete calculations in terms of expected value, it seems safe to estimate that giving A has the best expected consequences. Consequently, according to MSAC, it is right for Jill to administer A. At the same time, Jill does *know* that administering A does not have the best consequences. After all, only B or C can cure John. As soon as we accept that the subjective, epistemic situatedness of an agent plays a role with regard to the question of what is morally right for them, this judgment also seems to be the intuitively correct one. Taking a gamble and freely choosing between B and C, however, seems morally condemnable. Even if Jill were lucky enough to choose the

[76] Parfit made a similar point in an unpublished piece (with a case sometimes called The Miners; Parfit 1988) some years earlier. While it is plausible to assume that Parfit's case motivated Jackson's, a very similar case can be found in Donald Regan's book. However, Regan's example is hidden in a footnote at the very end of a book on the CHALLENGE (Regan 1980, pp. 264–265, footnote 1 of Chapter 11) where he also already concludes that "[in] case the reader is worried by the fact that my practical suggestions sometimes lead agents to abandon the attempt to produce the best consequences theoretically possible, I note that the same is true of a sensible practical approach to the application of any consequentialist theory." I add all this just to emphasize that considerations of this kind are quite close to the CHALLENGE: both Parfit 1988 and even more so Regan 1980 are important sources with respect to the CHALLENGE and will enter the stage again and again.

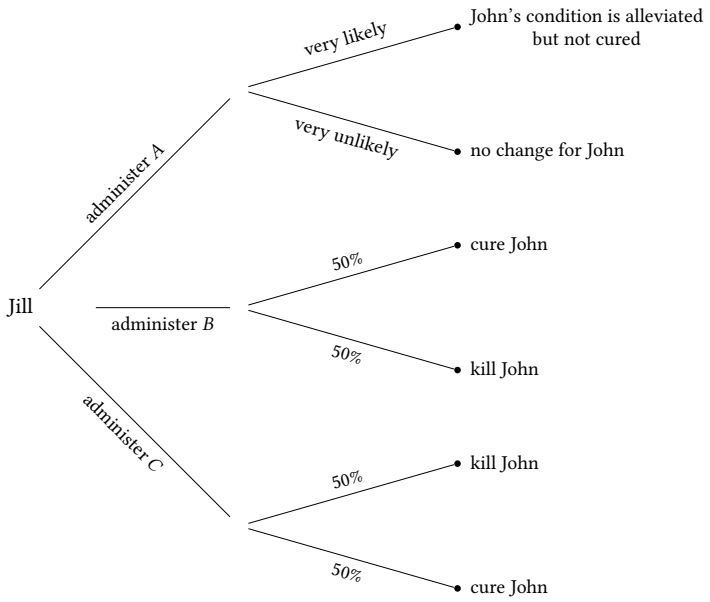


Figure 3.13: The extensive form of The Drug.

curative drug, it would seem appropriate to reproach her morally. To play a fifty-fifty cure-vs.-death lottery in the presence of a minor albeit non-trivial harm seems to be negligent, especially in the presence of an option that will very likely alleviate the condition without the risk of causing additional harm.[77]

Examples like The Drug offer many valuable lessons. However, the most essential takeaway regarding the Challenge is that subjective theories are generally willing to accept situations where doing the right thing even *excludes the possibility* of bringing about the morally best outcome. While The Drug demonstrates this in the context of individual decision situations, there is no obvious reason why the same lesson shouldn't apply to collective decision situations. Once a theory consciously rejects Congruence, it arguably might, *ceteris paribus*, also dismiss the two readings of it.

Certainly, Pinkert's formulation of On-the-hook seems rather convincing at first glance since it excludes precisely those circumstances that make guaranteed suboptimality morally tolerable, namely "non-culpable misinformation or lack of information". (Note that we previously encountered this general idea in connection with Epistemic Limes and Broad's Property, as discussed in Subsection 3.5.1.)

[77] A Subjective View like Railton's (1984), according to which, very roughly, one ought to aim for the best outcome, apparently would recommend either to administer B or C, but definitely not to administer A (as this excludes the best for sure). As this seems rather implausible, it seems quite reasonable that decision-theoretic, prospective subjective theories have prevailed with in the subjectivist's camp.

At this point, I do not want to take a position on whether Pinkert's claim is ultimately convincing. Even if it isn't, a variant of the No-Difference Challenge certainly remains for msac theories (cf. Hedden 2020; Kagan 2011; Singer 1980). Although msac was explicitly excluded from the scope of this thesis, the following observation remains relevant: there are qualified formulations of mh that could prove attractive, even for non-objective consequentialist theories. Thus, if it turns out (as it will!) that only a toned-down variant is defensible for moac, this would be acceptable—provided there are sound reasons for adopting such a modification. To borrow from Pinkert's formulation: a moral theory that accommodates Congruence better is, other things being equal, strongly preferable to a moral theory that accommodates it less well.

Even though every author has put the idea slightly differently, all of them apparently try to capture the same basic intuition, and "Principle of Moral Harmony", coined by Feldman, remains the most widely accepted name for it. Accordingly, I use "pmh" as a name for the general idea that connects all the specific phrases. In contrast, I use "Collectively Maximizing" for reference to the property and "mh" to denote the specific criterion which I introduced in Chapter 1 and which I will use in the context of this work. Recall

Criterion 1.2 (Moral Harmony (mh) – tentative) *A moral theory is adequate only if it is true that if all agents act rightly (according to this theory), then they are guaranteed to produce the morally best outcome they could bring about together.*

Note that, unlike Portmore (2018, p. 13), I propose that pmh should be understood as a necessary, but not a necessary *and* sufficient, criterion of adequacy. Otherwise, one would be committed to accept (in that sense) trivially adequate yet widely rejected moral theories and might also be forced into an untenable form of moral pluralism.[78]

[78] Assume we were to accept Portmore's version. Recall

Criterion 3.1 (mh– Portmore) *A moral theory T is correct if and only if the agents who satisfy T, whoever and however numerous they may be, are guaranteed to produce the morally best world that they could together bring about.*

There is a reading of this criterion that boils down to

(9) A moral theory is correct if and only if [if all agents act right according to that theory, then, necessarily, they produce the morally best outcome that they could together bring about].

According to such a criterion, every theory would be correct that makes true the conditional on the right-hand side of (9), i.e.,

(10) If all agents act right according to that theory, then, necessarily, they produce the morally best outcome that they could together bring about,

To argue in favor of the truth of P_{MOCOR}, the formulation of MH is sufficient. To see this, we rewrite it[79] to:

Criterion 1.3 (Moral Harmony (MH) – tentative, contraposition) *If a moral theory is adequate, then, if the agents in a collective decision situation produce a morally suboptimal outcome, (necessarily) at least one of the agents acted wrongly (according to this theory).*

This, in turn, arguably yields:

(P_{MH}) If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (according to this theory).

The preceding textual work, mind you, has not put forward any further substantive arguments for PMH beyond the justification already formulated in the introduction that Congruence allows for both an individual reading (most notably, MOCOR) and a collective one (PMH). However, it *should* have shown that the intuitive relevance and persuasiveness of PMH were and still are shared in large parts of moral-philosophical discourse. This should be enough for us to accept the premise as settled for now, even though, regarding

Now consider the following, rather prohibitive moral theory:

Definition 3.3 (Vacuous Criterion of Rightness (VacCor)) *All options are (necessarily) wrong.*

According to the standard Lewisian account, counterfactuals with impossible antecedences are vacuously true, in analogy to material conditionals with actually false antecedences. (In the second chapter of his *Counterfactuals* (1973), David Lewis accepts this view as a result of his semantic account but also briefly discusses alternative truth conditions. For a more detailed assessment of different approaches to this problem of 'counterpossibles' and their pros and cons, see D. H. Cohen 1987 and more recently Ferreira 2018. Lewis' standard account certainly suffices to motivate the issue in the given context.) Under this account, (10) is vacuously true and, thus, any theory embracing VacCor is guaranteed to fulfill MH and, therefore, to be correct according to (9) simply because it is then impossible for agents to perform the right actions. This seems plain wrong.

In addition, several theories might fulfill (10), most likely committing a champion of (9) to some implausible pluralism with respect to rightness.

[79] One merely takes the contraposition of the right-hand side of Criterion 1.2 and (arguably plausibly) reads the "guaranteed" as indicating a strict conditional. Furthermore, I translate the logically guaranteed "not act rightly" directly into "act wrongly" (but keep in mind the two possible MOAC wrongness predicates discussed at the end of Chapter 2—I come back to this later). As indicated in footnote 8, this translation is actually less innocent than one might initially think in light of the Consequentialist Standard View. However, it certainly fits Pinkert's On-the-hook and can be justified (which I will do later in Part 2 when the semantic difference actually matters).

its deeper semantic structure, the seemingly straightforward statement of MH is more complex than it initially appears. These nuances will be explored later in this thesis, where working them out in a precise and concise way will prepare the ground for the second, reconstructive part of my project. For now, however, we join the obviously widespread view elaborated here and next look at how one can derive from the existence of Troublemakers, via MOCOR (or via similar explications of MOAC's criterion of rightness) and against the background of PMH, an alleged inner-theoretical inconsistency.

3.5.2.3 Step 3: Regan's 'Proof' and P_{MOCOR}

The second of the "two particular intuitions [that] cannot be reconciled" (see above)—and thus the second key component of Regan's main argument—is a property named after act-utilitarianism (Regan 1980, pp. 3–4, 6):

Property 3.3 (PropAU (Regan)) *For any agent, in any choice situation, if the agent satisfies T in that situation, he produces by his act the best consequences he can possibly produce in that situation.*

From PropAU, in combination with PropCOP, Regan derives a third property, which he calls (ibid., p. 6):

Property 3.4 (Adaptability (Regan)) *A theory T is adaptable if and only if the agents who satisfy T, whoever and however numerous they may be, are guaranteed to produce the best consequences possible as a group, given the behavior of everyone else.*

Regan asserts that Adaptability is a "generalization of PropAU and PropCOP" that "is stronger than the conjunction of PropAU and PropCOP" (ibid., pp. 6–7). He supports this claim by proving that Adaptability implies both PropAU as well as PropCOP and by providing an intuition-based argument to demonstrate its greater strength, beyond mere equivalence to their conjunction (cf. ibid., pp. 107–108). For Regan, Adaptability is the paramount desideratum of exclusively act-oriented theories in general and, hence, of objective consequentialist theory more specifically. Finally, Regan's ultimate goal is to establish an impossibility result, namely that *no exclusively act-oriented theory can be adaptable.*

Regan does so by allegedly showing that there are cases—like Whiff and Poof and Two Factories, i.e., Mutual Exculpation Cases—in which actions that are right according to such a theory lead to a violation of PropCOP (and, thus, of MH), i.e., to suboptimal outcomes. Establishing such assessments is, thus, the third and final building block of Regan's argument and, hence, of the Challenge as the Internal Challenge.

Some authors apparently consider such assessments so trivial that they choose not to provide explicit arguments for them. For instance, recall that Pinkert claimed that "Act Consequentialism judges that both Ann and Ben act rightly by polluting" (Pinkert 2015, p. 975), which is based simply on the apparent observation that "no one could have brought about better results by acting differently" (ibid., p. 972). But the (often implicit) argument behind these inferences is actually quite interesting and informative when put under the microscope. Regan is explicit in this regard.

He starts his reasoning from his "precise necessary condition for exclusive act-orientation" (Regan 1980, p. 114). Recall his "partial definition":

> Any exclusively act-oriented theory must, in this example, on any assumption about Poof's (Whiff's) behavior, identify some non-empty subset of the set of acts comprising "pushing" and "not-pushing" such that Whiff (Poof) satisfies the theory if and only if he does some act from that subset. This necessary condition for a theory's being exclusively act-oriented I shall refer to, for expository convenience, as the 'partial definition' of exclusive act-orientation.

Closer inspection reveals that there are *two* characteristics of exclusive act-oriented theories. First, for an exclusively act-oriented theory, what makes an action right or wrong is *solely* based on features of the action itself (such as its consequences, compliance with specific rules, etc.), irrespective of, for instance, *why* the agent performed the action or what kind of person the agent is. Thus, examples of theories that are *not* exclusively act-oriented are all kinds of motive-based theories, such as virtue ethics or some versions of deontological ethics, which argue that the morality of an action is, at least in part, a function of the agent's motives, intentions, or character traits (cf. Section 2.2).

Concerning this first characteristic, it is rather obvious that MOAC theories are exclusively act-oriented (or, at least, fulfill Regan's necessary condition) by definition. Recall the formal definition of objective consequentialist theories:

Definition 2.3 (Objective Consequentialist Theory (formal)) *T is an* objective consequentialist theory *if and only if it embraces an axiological subtheory T_{Ax} with a valuation function $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$ and an objective consequentialist criterion of rightness T_{CoR} such that, for all decision situations $D \in \mathfrak{I}$ and for all $\phi \in \Phi_D$: $D, C \vDash_T \mathrm{R}\phi$ if and only if $T_{\mathrm{CoR}}(\phi)$.*

A criterion of rightness T_{CoR} is objective consequentialist *if and only if, for all $D \in \mathfrak{I}$ with $\mathsf{D} := \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O} \rangle$ (with D's actual context C) T_{CoR} corresponds to a predicate $\chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}$ such that for all $\phi \in \Phi$:*

$$D, C \vDash_T \mathrm{R}\phi \quad \textit{if and only if} \quad \chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}(\mathsf{Val}(\mathsf{Out}_C(\phi))).$$

Further, recall that MOAC theories are defined as *objective*, since their characteristic function can be explicated as

$$\chi^{\mathsf{Val}}_{\text{MOCOR},D,C}(\mathsf{Val}(\mathsf{Out}_{D,C}(\phi))) = \begin{cases} \top & \text{if } \phi \in \arg\max_{\phi' \in \Phi_D} \mathsf{Val}(\mathsf{Out}_{D,C}(\phi')), \\ \bot & \text{otherwise} \end{cases}$$

MOCOR's characteristic function obviously operates solely on the features of the options (more precisely on their consequences) and, thus, fulfills this first condition of Regan's partial definition.

Second, and easily overlooked, Regan's definition presupposes that the 'identified subsets'—representing permissible or right actions as defined by the theory under consideration – are *non-empty*. This just means that for every decision situation, there exists at least one morally right option.[80] We should articulate this second characteristic as a specific property of moral theories:[81]

Property 3.5 (Resolvability) *A moral theory is* resolvable *if and only if, for all decision situations and all relevant contexts, at least one option is right.*

Resolvability entails two other pertinent properties:[82]

Property 3.6 (Weak Deontic Completeness) *A moral theory is* weakly deontically complete *if and only if, for all decision situations and all relevant contexts, at least one action has a deontic status.*

and

Property 3.7 (No Moral Dilemmas) *A moral theory is* free of moral dilemmas *if and only if, for all decision situations and all relevant contexts, not all actions are wrong.*

Given that Resolvability inherently implies both Weak Deontic Completeness[83] and No Moral Dilemmas, it's apt to term it as being *stronger* than the

[80] Note that this assumption prevents, for instance, VACCOR from qualifying as being exclusively act-oriented (cf. footnote 78).

[81] I assume that Regan implicitly presupposes this for arbitrary decision situations, meaning he assumes that $\mathrm{I}_T = \mathrm{I}_T$.

[82] Similar properties have been considered for decision-making procedures in general, most importantly in the context of potential voids of responsibility, cf. Braham and Hees 2011.

[83] The term "weak" in Weak Deontic Completeness is used to differentiate it most importantly from:

Property 3.8 (Deontic Completeness) *A moral theory is* deontically complete *if and only if, for all decision situations and all relevant contexts, all actions have a deontic status.*

This property is stronger than Weak Deontic Completeness but isn't entailed by Resolvability.

two.[84]

At first glance, all these properties seem quite typical of consequentialist theories—and later, they indeed play a central role as elements of a set of criteria for assessing proposed solutions to the CHALLENGE. (But I am getting ahead of myself here.) If, for example, MOCOR tells us that exactly those actions are right and have at least as good consequences as all their alternatives, it may seem clear that *at least* one action must have this property.[85]

The crucial question now is: what options does this non-empty subset include in Mutual Exculpation Cases such as Whiff and Poof and Two Factories, according to MOAC (or, in Regan's terms, PropAU)? This question is far from trivial, as closer inspection reveals that it is underspecified *which* actions yield the best consequences and, therefore, 'satisfy' MOAC. To see this, recall the normal form of Whiff and Poof:

		Poof	
		not-push	push
Whiff	not-push	6	0
	push	0	10

[84]This use of "stronger" is consistent with the conventional understanding in formal logic. Specifically, a proposition α is deemed stronger than another proposition β if the truth of α restricts the set of potential models more than β, i.e., the set of models satisfying α is a strict subset of the set of models satisfying β. Essentially, α is true only in a narrower set of circumstances than β. This means α entails β without the converse being true. By extension, within the domain of moral theories and their properties, a property F is stronger than another property G when any moral theory satisfying F also satisfies G, but not (necessarily) vice versa.

Similar thoughts justify calling the two intertwined strands of the CHALLENGE, i.e., the INTERNAL CHALLENGE and the NO-DIFFERENCE CHALLENGE, the *stronger* and the *weaker* strand, respectively. Their interrelation is anchored in the hierarchy of their foundational assumptions. The INTERNAL CHALLENGE is built on a set of assumptions that are not only more rigorous but also 'encompass' the assumptions of the NO-DIFFERENCE CHALLENGE —if not strictly in a logical or conceptual sense, then at least intuitively. A MOAC theory that violates MH would arguably be counter-intuitive, at least from a consequentialist perspective (a claim I trust is adequately supported by my literature review at this point). This means that any theory satisfying the assumptions of the INTERNAL CHALLENGE will, *arguably*, satisfy those of the NO-DIFFERENCE CHALLENGE. Therefore, if the challenge presented by the INTERNAL CHALLENGE is substantiated, they represent a more formidable and encompassing critique of consequentialism. Given this logical entailment and the differential implications for consequentialism, it's fitting to term the INTERNAL CHALLENGE as the stronger strand and the NO-DIFFERENCE CHALLENGE as the weaker strand. This terminology aids in demarcating the two while also highlighting the inherent intensity and breadth of challenges they pose for MOAC theories.

[85]Properties along these lines have been proposed as distinguishing consequentialist from non-consequentialist but 'consequentializable' theories (cf. Dreier 1993, 2011; Portmore 2009); see also Brown 2011 in relation to Property 3.7.

Let's ask what options of Whiff (and Poof, respectively) in Whiff and Poof are right according to MOAC. This is to ask, for $X \in \{\text{Whiff}, \text{Poof}\}$, which non-empty subset S_X of

$$\Phi_X := \{\, \text{pushing}_X, \text{not-pushing}_X\}$$

corresponds to the characteristic function $\chi^{\mathsf{Val}}_{\text{MOCOR},D,C}$. The right answer, it seems, is: it depends on what the respective other agent does. And this is what Regan's central argument 'exploits'.

Here is Regan, unfolding his 'impossibility result' (Regan 1980, p. 115):

> Suppose there is an adaptable theory T which satisfies the partial definition. Suppose further that Poof does not push. Since T satisfies the partial definition, there is some non-empty subset of the set of acts "pushing" and "not-pushing" such that Whiff satisfies T (while Poof does not push) if and only if he does an act from that subset. Call the subset S. We can deduce what S must be from the assumptions we have made about T. We know that Whiff satisfies T if and only if he does an act from S. So, if Whiff does an act from S, he satisfies T. Since T is adaptable, T has PropAU. That means that any agent who satisfies T produces the best possible consequences in his circumstances. If Whiff produces best possible consequences in his circumstances, which include Poof's not-pushing, he must not-push. Therefore, if Whiff satisfies T, he not-pushes. Remembering what we have already established, that if Whiff does an act from S, he satisfies T, we can conclude that if Whiff does an act from S, he not-pushes. But remember also that S is non-empty. The only non-empty set such that if Whiff does an act from that set he not-pushes is of course the set consisting of the act "not-pushing". Therefore S consists of the act "not-pushing". In sum, if Poof does not push, then Whiff satisfies T if and only if he (Whiff) not-pushes also.

We can reconstruct Regan's argument in terms of the formalism introduced in Chapter 2: First, we assume that there is a theory T that is adaptable and satisfies Regan's partial definition of exclusive act-orientation. From this, we infer for Whiff, relative to D representing Whiff's individual decision situation in[86] Whiff and Poof with its actual context C:

1. Since T fulfills Regan's partial definition, there is a set $S \subseteq \Phi_{\mathsf{D}}$ such that
   a) $|S| > 0$, i.e., $S \neq \emptyset$ and
   b) for all $\phi \in \Phi_{\mathsf{D}}$: $T, \mathsf{D}, C \models \phi$ if and only if $\phi \in S$.

[86] The implicit assumption that there is such an individual decision fits well with our corresponding presumption, cf. 46 in Section 3.3.

2. Since T is adaptable, T entails PropAU—which we assume to be extensionally equivalent to MOCOR—and, thus, for all $\phi \in \Phi_{\mathsf{D}}$, if $T, \mathsf{D}, C \vDash \phi$, then $\phi \in \arg\max_{\phi' \in \Phi_{\mathsf{D}}} \mathsf{Val}(\mathsf{Out}_{\mathsf{D},C}(\phi'))$.[87]

Thus, unsurprisingly, in light of the formerly established formalism, it holds that

$$S = \arg\max_{\phi' \in \Phi_{\mathsf{D}}} \mathsf{Val}(\mathsf{Out}_{\mathsf{D},C}(\phi')).$$

Regan claims that this is sufficient to derive what is right for Whiff (according to such a theory T) *given that Poof not-pushes*. To assume that Poof not-pushes is to assume that C includes that Poof not-pushes. In this case, we know that Whiff produces the best possible consequences only if he not-pushes. Thus, we know that

$$S = \arg\max_{\phi \in \Phi_{\mathsf{D}}} \mathsf{Val}(\mathsf{Out}_{\mathsf{D},C}(\phi)) = \{\text{not-pushing}_{\mathsf{W}}\}.$$

This, of course, is simply to say that Whiff not-pushing would have better consequences than Whiff pushing. Thus we know that, according to 2.,

(11) If $T, \mathsf{D}, C \vDash \phi_{\mathsf{D}}$, then $\phi_{\mathsf{D}} = \text{not-pushing}_{\mathsf{D}}$,

that is, as Regan correctly states, to say that, relative to C

(12) If Whiff satisfies T, then he not-pushes.

Equivalently, it holds, again relative to C, that:

(13) If Whiff does an act from S, he not-pushes.

Thus, either there is no act in S or, if there is, then it is not-pushing. But (1.a.) warrants that there is at least one act in S. Thus it is right for Whiff to not-push.

It is straightforward to demonstrate the same result with the roles reversed (Regan 1980, p. 115):

> What we have just proved about Whiff we could also prove about Poof. Given our assumptions about T, if Whiff does not push, then Poof satisfies T if and only if he (Poof) does not push.

This leads Regan to the conclusion that if both not-push, they both do right according to T (ibid., p. 116):

[87] This is exactly one direction of the condition we have defined for the predicate of MOCOR.

> But now, suppose that both Whiff and Poof not-push. We have demonstrated that each of them satisfies T, when the other not-pushes, if and only if he not-pushes. Therefore, when both not-push, both satisfy T.

The next step of Regan's argument cannot come as a surprise to anyone (ibid., p. 116):

> But then universal satisfaction of T does not guarantee the production of the best possible consequences. T does not have PropCOP. Adaptability entails PropCOP, so if T does not have PropCOP, T is not adaptable. That is a contradiction. We conclude that a theory which satisifies [sic!] the partial definition of exclusive act-orientation cannot be adaptable. QED.

We can now take a step back and try to get a better grip on what actually happens in Regan's argument. For this, it is crucial to emphasize how strongly Regan's reasoning relies on the idea that the moral status of what is right for Whiff and Poof must be carefully assessed relative to the *exact* context. Take, for instance,

(14) It is right for Whiff not to push.

Is (14) true *relative* to the context defined solely by Whiff and Poof (i.e., without making further assumptions about their decisions and actions), or is it false? It seems that this is a false dichotomy. Instead, it seems plausible to understand it as underspecified or as undefined, i.e., as neither true nor false.[88] However, relative to Whiff and Poof, *together* with the fact that

Fact 3.4 *Poof does not push*

it appears that (14) *is* true. At least, this is what Regan's argument above apparently warrants.

Does this mean that we cannot say *anything* about what's right in Whiff and Poof? Not quite, because such considerations allow the derivation of *conditional* assessments. For instance, it seems plausible that

(15) If Poof does not push, then it is right for Whiff not to push

is true relative to Whiff and Poof.

Note that, based on the considerations presented thus far, it remains unsettled whether such conditional assessments are best understood as indicating that the options of Whiff and Poof possess, relative to Whiff and Poof (i.e., without additional assumptions about the other agent's actions), a *genuine*

[88] This is not particularly spectacular. Jan Łukasiewicz, C. I. Lewis, and others have drawn their motivation for the development of three-valued logic from insufficiently specified propositions (cf. C. I. Lewis 1932; McCall 1973.

conditional deontic status or instead *no* deontic status at all. While I want to emphasize that both understandings are at odds with what I called the Consequentialist Standard View (cf. Section 2.2, page 21), we do not need to take a stand on that question for now (but I will revisit the question later).

It is useful to introduce a compact notation for the concepts used here. Let D represent a collective decision situation, C the actual context, F a state of affairs, and f the proposition that F obtains. For the purposes of this project, F generally pertains to what some agents in D do, will do, or have done. To abbreviate, we define "$[\![C \oplus F]\!]$" as short for "C together with the assumption that F obtains". In essence, this means that the actual context (or "the circumstances", as Regan would say) of the decision situation D is *extended* by F. This allows us to articulate the technique of *conditionalization* in the form of the following principle:

Principle 3.10 (Conditionalization) *Let C be a context, and let F be some state of affairs. If it is true, relative to $[\![C \oplus F]\!]$, that p, then it is true, relative to context C, that [if f, then p].*

Conditionalization suggests an inference scheme. For a (collective) decision situation D with context C and a state of affairs F and a propositional variable φ,

$$T, D, [\![C \oplus F]\!] \vDash \varphi \quad \text{implies that} \quad T, D, C \vDash f \to \varphi.$$

Precisely speaking, the first half of Regan's argument has thus established nothing but the truth of

(16) Whiff does right if and only if Whiff does also not push.

relative to $[\![\text{Whiff and Poof} \oplus \text{Fact 3.4}]\!]$ (and T). Therefore, according to Principle 3.10, he has also established that

(17) If Poof does not push, then Whiff does right if and only if Whiff does also not push

relative to Whiff and Poof (and T).

In the same way, we can certainly introduce

Fact 3.5 *Whiff does not push.*

and derive that it is true that

(18) Poof does right if and only if Poof does also not push

relative to $[\![\text{Whiff and Poof} \oplus \text{Fact 3.5}]\!]$ (and T). Therefore, according to Principle 3.10 he also established that

(19) If Whiff does not push, then Poof does right if and only if Poof does also not push

relative to Whiff and Poof (and T).

As soon as (17) and (19) have been established as true (relative to Whiff and Poof and T), Regan's argument proceeds with the consideration of

Fact 3.6 *Poof does not push and Whiff does not push.*

Regan now claims that, from (17) and (19), we can infer

(20) Poof does right and Whiff does right

relative to [[Whiff and Poof $\oplus$ Fact 3.6]] (and T).

It is, then, tempting to read Regan's argument simply like this:

Licensing Not-Pushing with Regan (naïve)

$P^{\text{Regan}}_{P\to R(W)}$: If Poof does not push, then Whiff does right if and only if Whiff does not push.

$P^{\text{Regan}}_{W\to R(P)}$: If Whiff does not push, then Poof does right if and only if Poof does not push.

$P^{\text{Regan}}_{W\wedge P}$: Whiff does not push and Poof does not push.

$C^{\text{Regan}}_{R(W)\wedge R(P)}$: Poof does right and Whiff does right.

This reconstruction, however, might be guilty of glossing over important semantic aspects of Regan's argument. For instance, one could (and maybe should) ask what the *frame of reference* of this argument is supposed to be: the first two premises are established relative to Whiff and Poof, but the third premise is *false* relative to Whiff and Poof and only true relative to [[Whiff and Poof $\oplus$ Fact 3.6]]. For [[Whiff and Poof $\oplus$ Fact 3.6]], however, we have not established the first two premises. This makes it hard to properly assess this naïve reconstruction semantically.

An alternative and probably better way of putting Regan's argument, then, is to reconstruct it in terms of some kind of metalanguage:

Licensing Not-Pushing with Regan (meta)

$P^{\text{Regan, meta}}_{P\to R(W)}$: In Whiff and Poof: if Poof does not push, then Whiff does right if and only if Whiff does not push.

$P^{\text{Regan, meta}}_{W \to R(P)}$: In Whiff and Poof: if Whiff does not push, then Poof does right if and only if Poof does not push.

$C^{\text{Regan, meta}}_{R(W) \wedge R(P)}$: In ⟦Whiff and Poof ⊕ Fact 3.6⟧: Poof does right and Whiff does right.

I think this way of stating the argument is proper, but it seems as if some bridging principle is missing that would allow us to combine and extend contexts in the required way. However, to the best of my knowledge, Regan's reasoning has never been challenged. On the contrary, as we will see below, it is repeated in a similar form to this day. Therefore, in the spirit of reconstruction, I will devote myself briefly to those similar argumentations and postpone a more detailed examination of the above argument with respect to its soundness until the second part of this thesis.

In the literature on the Challenge, the first reconstruction is quite common. Recall David Estlund and his Trilemma from the beginning of this chapter. He described a structurally similar case (Estlund 2017, p. 53):

Case 3.9 (Dr. Slice and Dr. Patch) *Dr. Slice is a surgeon and his colleague Dr. Patch is an expert in stitching up wounds. They are faced with a situation where a patient has a tumor. If Dr. Slice makes an incision to remove the tumor, it is necessary that Dr. Patch (or someone else) stitches up the wound afterward. If the patient is both cut and stitched, his life will be saved. However, if there is surgery without stitching or (for whatever crazy reasons) stitching without surgery, the patient will have an agonizing death. If nothing happens, the patient will die but will be spared some pain.*

Here is the corresponding decision situation in normal form:

		Patch	
		go golfing	patch
Slice	go golfing	the patient dies	the patient dies and suffers
	cut	the patient dies and suffers	the patient survives

Now, as the story goes, on this particular occasion, both Dr. Slice and Dr. Patch have plans to go golfing. Dr. Slice will not perform the surgery and Dr. Patch will not be available to stitch up the wound. As a result, the patient's condition worsens and he eventually dies. It is easy to see that Dr. Slice and Dr. Patch is a Troublemaker and that the actually instantiated combination of actions structurally corresponds to Whiff and Poof both not-pushing their buttons. Here is Estlund's moral assessment of that particular situation (ibid., p. 53):

> What Slice is required to do depends on what Patch will do. [...]
>
> Patch ought to stitch the patient if and only if Slice will be doing the surgery (stitching is possible, but pointless and harmful if there is no wound that needs stitching). But suppose that Slice will not be doing the surgery. Patch might as well go golfing. Ought Slice to cut? Well, no, because Patch will not be there to stitch, and so the surgery will only make the patient's death more painful. Slice might as well go golfing. Neither has acted (or omitted) wrongly, despite the fact that the patient will needlessly die.

Translating a bit from prescriptive to rather descriptive deontic vocabulary, we can reconstruct Estlund's argument very much like Regan's:

Licensing Golfing with Estlund

$P^{\text{Estlund}}_{R(P)\leftrightarrow S}$: It is right for Patch to stitch the patient if and only if Slice will be doing the surgery.

$P^{\text{Estlund}}_{R(S)\leftrightarrow P}$: It is right for Slice to do the surgery if and only if Patch will be stitching.

$P^{\text{Estlund}}_{\neg S\wedge\neg P}$: Slice will not be doing the surgery and Patch will not be stitching.

$C^{\text{Estlund}}_{R(P)\wedge R(S)}$: It is right for Slice to go golfing and it is right for Stitch to go golfing.

Besides the minor difference in grammatical tense, I think it is fair to say that Eslund essentially mimics Regan's argument. The only difference seems to be the temporal perspective. Where Regan uses the present tense, Estlund considers, in simple future, what the agents *will* do.

Finally, consider Felix Pinkert, the inventor of Two Factories. As mentioned earlier, we read (Pinkert 2015, pp. 974–975):

> [Two Factories] becomes a challenge for Act Consequentialism only once we assume that Ann and Ben are both 'uncooperative,' that is, each would pollute even if the other produced cleanly. [...] In The Two Factories, it is only if both agents are uncooperative that neither could have improved matters by acting differently and that Act Consequentialism judges that both act rightly. [...]
>
> Ann and Ben each individually could only have made matters worse by producing cleanly, as the other agent would then still have polluted the river, and the livelihoods of 100 workers in the cleanly producing factory would have been destroyed. [...]
>
> Act Consequentialism judges that both Ann and Ben act rightly by polluting, even though they together could easily have brought about much

> better outcomes by both producing cleanly. So if Consequentialists only have Act Consequentialism to morally apprise [sic!] a situation, then they let Ann and Ben off the hook for together producing collectively suboptimal outcomes.

We can reconstruct Pinkert's argument in different ways, but if we look for a difference relative to the former two, we find one reading that is based on a backward-looking, post-hoc description. What matters in Pinkert's reasoning is the fact that they "could only have made matters worse" by acting otherwise. We might go for this reconstruction:

Licensing Polluting with Pinkert

$P^{\text{Pinkert}}_{A\to R(B)}$:	Given that Ann did pollute, Ben could only have made matters worse by producing cleanly.
$P^{\text{Pinkert}}_{B\to R(A)}$:	Given that Ben did pollute, Ann could only have made matters worse by producing cleanly.
$P^{\text{Pinkert}}_{A\wedge B}$:	Ann did pollute and Ben did pollute.
$P^{\text{Pinkert}}_{\text{BW}-\text{MOCOR}}$:	If an agent could only have made things worse by acting otherwise, then they did right.
$C^{\text{Pinkert}}_{R(A)\wedge R(B)}$:	Ann did right by polluting and Ben did right by polluting.

The backward-looking character of Pinkert's argument distinguishes it from the arguments of Regan and Estlund. And the worries regarding the naïve reconstruction of Regan's argument do not apply here, as there is no shift in the frame of reference. Yet, the first three premises are warranted by reasoning very similar to that explicated above in the context of Regan's original argument, and $P^{\text{Pinkert}}_{\text{BW}-\text{MOCOR}}$ seems to be quite plausible, too, as it appears just like an innocent backward-looking lemma of MOCOR. We will later revisit Pinkert's argument and ask whether this temporal shift is really so innocent and whether it suffices to establish the fact that all agents actually *do* right by not pushing in Whiff and Poof.

One additional remark on Pinkert's reasoning is imperative. Pinkert correctly emphasizes the importance of the property that he calls uncooperativeness, which, as he points out, has been called *intransigence* by Michael J. Zimmerman (1996, p. 257). To see why it is important, consider

Fact 3.7 (Coop-Ben) *Actually, if Ann were to produce cleanly, Ben would produce cleanly (but if Ann were to pollute, Ben would pollute as well).*

Given Fact 3.7, we might say that Ben is willing to 'do his share' if others do as well. Such a 'tit-for-tat'-like disposition is not far-fetched, and we probably encounter corresponding attitudes quite often in everyday life. Given Ben's willingness to opt for the best outcome, however, Ann *could* bring about the best possible outcome by producing cleanly. In other words,

(21) It is right for Ann to produce cleanly and wrong for her to pollute

is true relative to ⟦Two Factories ⊕ Fact 3.7⟧ (excluding the fact that they actually act uncooperatively, obviously). Thus, the right and wrong thing to do sometimes does not only depend on what the other agent *does* but also on what they *would* do.

If we want to be pedantic, both terms—"(un)cooperativeness" and "intransigence"—are somewhat misleading in describing the property they refer to. The term "(un)cooperativeness" evokes ideas of communication, negotiation, shared goals, or joint payoffs—none of which are applicable in the cases under consideration. Similarly, the term "intransigence" captures a specific kind of behavior where agents stubbornly *insist* on their actions despite being informed of others' intentions. However, this only applies in certain subjective contexts, whereas we are dealing with an objective setting here.

This gives us a reason to adopt a more suitable term from game theory. For the remainder of this book, we will implicitly assume *independency of actions*—and we will also implicitly refer to this property by this label and presuppose it throughout: whatever agents do, they act independently of what other agents do.[89] This assumption is embedded in the description of Two Factories from the very beginning.

We can put this a bit more precisely. Let us to call[90] a combination of actions *proper* (relative to some collective decision situation) if (and only if) there is a consequence defined for it by the description of the case at hand. We can then define

Property 3.9 (Independency of Action) *Let $\mathcal{D}$ be a collective decision situation. A combination of actions that is proper within $\mathcal{D}$ is* act-independent *if and only if any combination resulting from a unilateral deviation from that original combination is also proper within $\mathcal{D}$. The decision situation $\mathcal{D}$ itself is called* act-independent *if and only if all combinations of actions in $\mathcal{D}$ are act-independent.*

[89] The notion corresponds to the typical assumption for *non*-cooperative games, cf. Nash 1951, p. 286: "Our theory, in contradistinction, is based on the absence of coalitions in that it is assumed that each participant acts independently, without collaboration or communication with any of the others."

[90] This rough-and-ready formulation will be explicated more precisely and formally later, but for now it does the job.

The idea here is that if all combinations resulting from a unilateral deviation from an arbitrary proper combination are also proper, this rules out the kind of dependency described in Fact 3.7, i.e., that an agent would alter their course of action based on the actions of another agent. Thus, let us assume for the remainder of this book that all cases under consideration are act-independent.[91]

Let me summarize the central takeaway message from this subsection. The first part of Regan's argument was meant to establish the last premise of what I called the ARGUMENT, viz.

> (P_{MOCOR}) If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

If Regan and his successors are correct, then the agents in Troublemakers, when performing the troublesome combinations, create, through their actions, for each other the contexts (or circumstances) sufficient to *mutually exculpate* each other's actions.

Regan's reasoning is detailed, and the passages and arguments collected and highlighted here emphasize that this aspect of Regan's conclusions resonates in contemporary philosophical discussions. This reaches far beyond the already cited examples. Derek Parfit (both in 1984 and 1988), Frank Jackson (1987), Michael J. Zimmerman (1996), and more recently, Shelly Kagan (2011) and Douglas Portmore (2018) have argued in similar terms, and many more examples can be found. That all agents do the right thing in the eyes of MOAC when realizing a troublesome combination of actions is more or less taken for granted in the literature.

[91] Strictly speaking, this assumption is stronger than necessary as it does not only exclude dependencies that would break the troublesomeness of *some* combinations of actions but *all* kinds of such dependencies. For instance, consider replacing the explicit independence assumption in the case description of Two Factories with the fact

Fact 3.8 (Contingent Anti-Ben) *Actually, if Ann were to produce cleanly, Ben would pollute (but if Ann were to pollute, Ben would decide independently).*

This dependency certainly would not break the troublesomeness of the combination that consists of both agents polluting, even though it would not align with Independency of Action (because, in light of Fact 3.8, the combination of both producing cleanly is not proper in the relevant sense even though Ann polluting and Ben producing cleanly is). But since these dependencies only add complexity to Troublemakers without raising new related challenges—besides, maybe, modeling issues—we can ignore them in the context of this investigation and proceed without further specifying or restricting Independency of Action.

To the best of my knowledge, the validity of reasoning in the tradition of Regan has never been seriously contested. However, Fred Feldman provides a passage that directly addresses a critical issue (Feldman 1980, p. 177; the original example has been modified to align with Regan's Whiff and Poof case):

> It is agreed that Whiff ought not to push if Poof does not push. It is also agreed that Poof does not push. However, from these two premises, we may not infer that Whiff absolutely ought not to push. We cannot detach an absolute obligation from a conditional obligation and its condition.

I think this is a *crucial* point (and will come back to this later). But Feldman thought that it is a minor one because he believed that, ultimately, we can detach the non-conditional obligation (ibid.):

> Yet in the case at hand I believe we have another premise that enables us to detach our conclusion. That premise is that the condition is 'inevitable.' More precisely, it is that no matter what Whiff does, Poof will not push. Nothing Whiff can do will make Poof push. This fact, together with the conditional obligation not to push if Poof does not, entails that Whiff absolutely ought not to push.

In other words, Feldman just highlighted the relevance of Independency of Action again.

Such easygoing and simple arguments have long convinced the philosophical community and should suffice for the purposes of my project—though we will revisit them later.

In the second part of his argument, then, Regan and his followers derived the inconsistency that threatens to haunt MOAC. That is, Regan's argument can indeed be straightforwardly translated into the Challenge as Internal Challenge in terms of the Argument. Recall:

The Argument – tentative

$P_{\exists\text{TROUBLE}}$: There are Troublemakers: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

P_{MOCOR}: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

P_{MH}: If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (according to this theory).

$C_{\neg\text{ADEQ}}$: MOAC is not an adequate moral theory.

Before I continue by turning to previous attempts to solve the CHALLENGE and what would characterize an acceptable solution in terms of this project, let's have a brief look back at the overall reconstruction of the CHALLENGE.

3.6 The Pyramid and the Next Steps

In the opening of this chapter, I introduced the concept of the CHALLENGE as not a single variant but a hierarchy of variants, which I dubbed the Pyramid (refer to Figure 3.14 for a recap). These variants of the CHALLENGE differ in severity and prerequisite-richness. Lower-ranked variants within the Pyramid are less severe, i.e., would have a less severe impact in case of success, but require weaker or fewer assumptions. Consequently, they have broader applicability, targeting not only objective but also subjective consequentialist theories or challenging even moral theories in general. Thus, lower-ranked variants can back up the higher ones in the Pyramid.

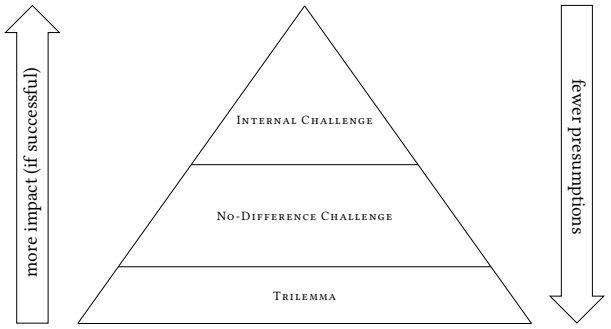


Figure 3.14: The three variants that make up the Pyramid as tackled in this thesis, ordered by their strength and dependence on the preconditions they presuppose.

At the apex of the Pyramid, we find the CHALLENGE as INTERNAL CHALLENGE, Regan's 'impossibility result'. Should this variant be validated, MOAC must be deemed to be internally inconsistent. This is contingent on the validity of MH. Next in line is the CHALLENGE as the NO-DIFFERENCE CHALLENGE, which, if successful, only convicts MOAC of extensional inadequacy. This version does not rely on the MH assumption and is relevant to all Difference-Making Views. At the foundation of the Pyramid, we find the highly generalized TRILEMMA.

I hope to have elucidated the claim that the Pyramid, compared to individual variants present in the literature, is a stronger manifestation of the

CHALLENGE. I have striven for precise representation of and sound anchoring within the literature. A comprehensive and satisfying resolution to the CHALLENGE, from the perspective of MOAC, would solve all three variants while maintaining the theory's integrity.

Of course, this does not say much about what constitutes a *good* solution. In the following chapter, I suggest such criteria and briefly offer some reasons why there is no satisfactory solution to the CHALLENGE so far.

Chapter 4

The Good, the Bad, and the Ugly

In this chapter, my focus shifts to proposed solutions to the Challenge. I begin by introducing what I call *solution spaces*, explicit structures that allow for taxonomizing different approaches to the Challenge and tackling it systematically. This construct will serve as a sort of compass for navigating the remainder of this project, as it defines the potential paths to solutions of the Challenge. Furthermore, in the initial application of these solution spaces, I employ them to justify excluding Cumulative Effects Cases from the scope of this project.

Next, I turn to the question of what actually makes a solution a good one. For this, I lay the foundation for evaluating and assessing approaches to the Challenge by establishing a set of specific criteria. These serve as benchmarks for evaluating the merits of different approaches and will guide my assessment of proposed solutions, offering a basis for measuring their effectiveness within the consequentialist framework.

Subsequently, I will examine three carefully selected (particularly influential or especially intriguing) concrete approaches. After positioning them within the solution spaces and evaluating them against the introduced criteria, I then examine whether these proposals offer satisfactory solutions to the Challenge from the perspective of moac.

Not surprisingly, my investigation find that none of the solutions proposed so far is satisfactory—a new approach is needed. This will conclude the first part of my work, and we then move on to the second part, where I develop such an approach.

4.1 Mapping the Solution Space

Before delving into what it means to devise a *good* solution to the Challenge within the scope of this project, it's essential to clarify what constitutes *a* solution in the first place.

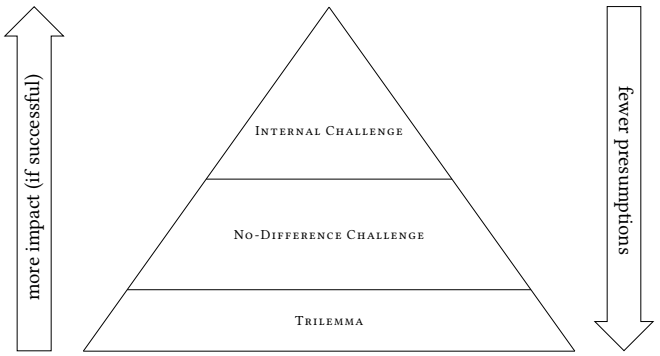


Figure 4.1: The three variants that comprise the Challenge as the Pyramid tackled in this thesis.

To do so, let's refer back to the previously established Pyramid (cf. Figure 4.1) and momentarily set aside any consequentialist constraints. I call a *solution* to the Challenge a theory that can master all these variants. A solution should start with discharging the most dangerous variant and then continue with the next. Hence, a solution can be imagined as a path that leads from top to bottom, i.e., *out* of the Challenge as Internal Challenge, safely *through* the Challenge as No-Difference Challenge, while simultaneously offering a way to *break out* of the Trilemma.

Thus, the first objective for a solution to the Challenge is to address its most perilous variant—the Challenge as Internal Challenge. This variant finds its expression in what I called the Argument. This variant finds its expression in what I called the Argument. As a reminder, here it is again:

The Argument – tentative

$P_{\exists\text{TROUBLE}}$: There are Troublemakers: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

P_{MOCOR}: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to moac).

P_{MH}: If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (according to this theory).

$C_{\neg\text{ADEQ}}$: moac is not an adequate moral theory.

One can respond to the ARGUMENT in two principal ways. First, one can try to *defend* the attacked theory. Second, typically, after failing with that first way, one can focus on *managing the defeat*. Here are more details from the specific perspective of MOAC:

1. **Defense:** If champions of MOAC want to defend against the ARGUMENT, they need to show that it is not sound. There are, as always, two lines of defense:

   Attack Validity: Show that something is *structurally* (and thus logically) wrong with the ARGUMENT and that it is, thus, invalid. This probably requires scrutinizing the logical-conceptual structure of the argument.

   Attack Truthfulness of a Premise: Show that something is *substantially* wrong with the ARGUMENT by rejecting one of the premises. So there are three points of leverage in this case:

   Refute $P_{\exists\text{TROUBLE}}$: Typically, $P_{\exists\text{TROUBLE}}$ is justified by presenting an example Troublemaker (or a class of Troublemakers inspired by one or more examples), which is usually defined through a specific description. So, to refute $P_{\exists\text{TROUBLE}}$, one has to show that there is something wrong with the description (or with a whole class of descriptions). There are two specific manifestations of this strategy:

   Challenging the Description: Raising doubt as to whether the supplied description satisfies the formerly introduced conditions, i.e., showing that there are reasons to believe that it is not coherent, consistent, or sufficiently complete (cf. Section 2.1, page 18).

   Challenging its Troublemaker-hood: Raising doubt as to whether a given description actually describes a Troublemaker, i.e., showing that there actually is no troublesome combination.

   Refute P_{MOCOR}: To refute P_{MOCOR}, one must show that the application of MOCOR identifies at least one wrong action within any troublesome combination.

   Refute P_{MH}: To refute P_{MH}, one must show that MH is no persuasive criterion of adequacy.

2. **Manage the Defeat:** If champions of MOAC fail to defend against the ARGUMENT along the above-sketched lines of defense, they can still try to make the best of the impending defeat. In principle, there are at least two paths to choose from:

Biting the Bullet: In principle, there is always the possibility that the side under attack may be prepared to live with the conclusion after all. However, with regard to the INTERNAL CHALLENGE, the option of bullet-biting can be ruled out for camp MOAC: accepting a theory while acknowledging its *inadequacy* derived from a diagnosed inconsistency within said theory itself appears to violate rational standards. No one should support a theory whose theoretical inconsistency they must admit.[92] To 'bite' $C_{\neg\text{Adeq}}$ as a result of the ARGUMENT would amount to just that for the followers of MOAC. Therefore, this approach cannot help with the INTERNAL CHALLENGE.

Modify & Adapt: Alternatively, one could try to modify one's theory to allow one to defend it against the ARGUMENT without incurring *new* problems. The aim is to keep the spirit of one's theory but change details, carry out extensions, or make conditionalizations. In the context of this project, of course, particular care must be taken to ensure that the modified theory remains a MOAC theory.

This leaves us with quite a number of alternatives for responding to the INTERNAL CHALLENGE. In Figure 4.2, we see a structured set of pathways—a comprehensive map I refer to as the "solution space of the ARGUMENT" (or of the INTERNAL CHALLENGE, respectively). Note that in the model, any theory reaching an endpoint marked as "successfully defended" can be considered a *solution* to the INTERNAL CHALLENGE (though not necessarily a satisfactory one, particularly from the fundamental consequentialist perspective adopted in my project; we will address this later in the chapter).

It is important to note that even after we have found a solution to the INTERNAL CHALLENGE, our work is not done. Next, we are confronted with the two other variants of the CHALLENGE, the two other levels of the Pyramid. The next hurdle, hence, is the NO-DIFFERENCE CHALLENGE. Recall that I also sketched an argument corresponding to that variant:

[92] Note that *supporting* a theory is not necessarily to believe in its adequacy. For instance, one may well be a supporter, in a sense, of a theory that one thinks to be *useful* even though one believes it to be inadequate in the sense of being false. For example, one might consider Newtonian mechanics a valuable theory for all use cases one considers relevant and support for it to be taught to students worldwide, even while being fully aware of general and special relativity. More importantly, every scientific theory is (at least very likely) false. We have indeed not reached the end of an imagined ideal scientific process, at the point of convergence of a Peirce-like theory of truth (cf. Peirce 1931)—and probably will never arrive there. But supporting a theory that one believes to be inadequate in the sense of being *inconsistent* is of a different caliber altogether. *Ex falso sequitur quodlibet* and the principle of explosion are no friends of serious theories.

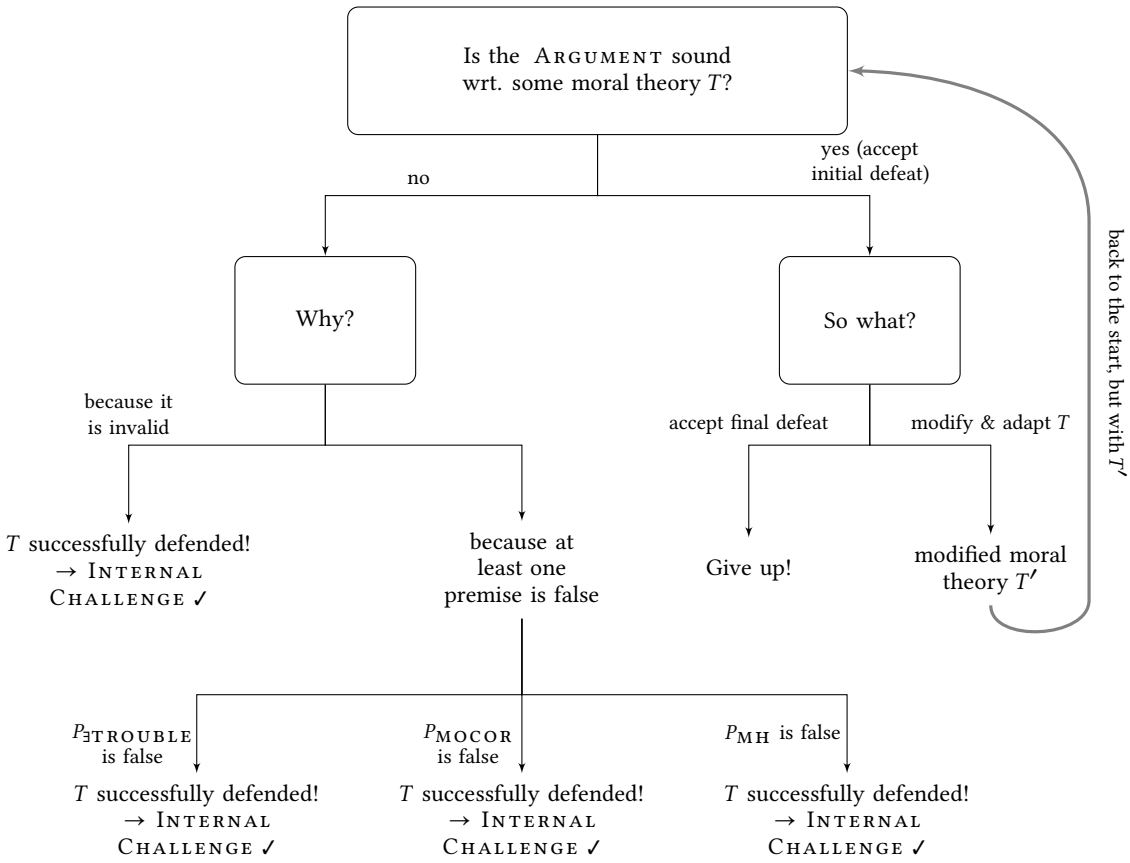


Figure 4.2: The solution space of the Challenge as Internal Challenge.

The No-Difference Challenge Argument – tentative

$P_{\exists \text{NDCs}}$: There are No-Difference Cases: collective decision situations in which there is at least one agent who can act in a way such that it seems intuitively morally wrong, but the agent could not make a difference for the morally better by unilaterally acting differently.

P_{MOCOR}: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

$P_{\neg\text{intu}}$: If a moral theory assesses intuitively morally wrong actions as morally right for a significant class of situations, then T is counterintuitive.

$C_{\neg\text{intu}}$: MOAC is counterintuitive.

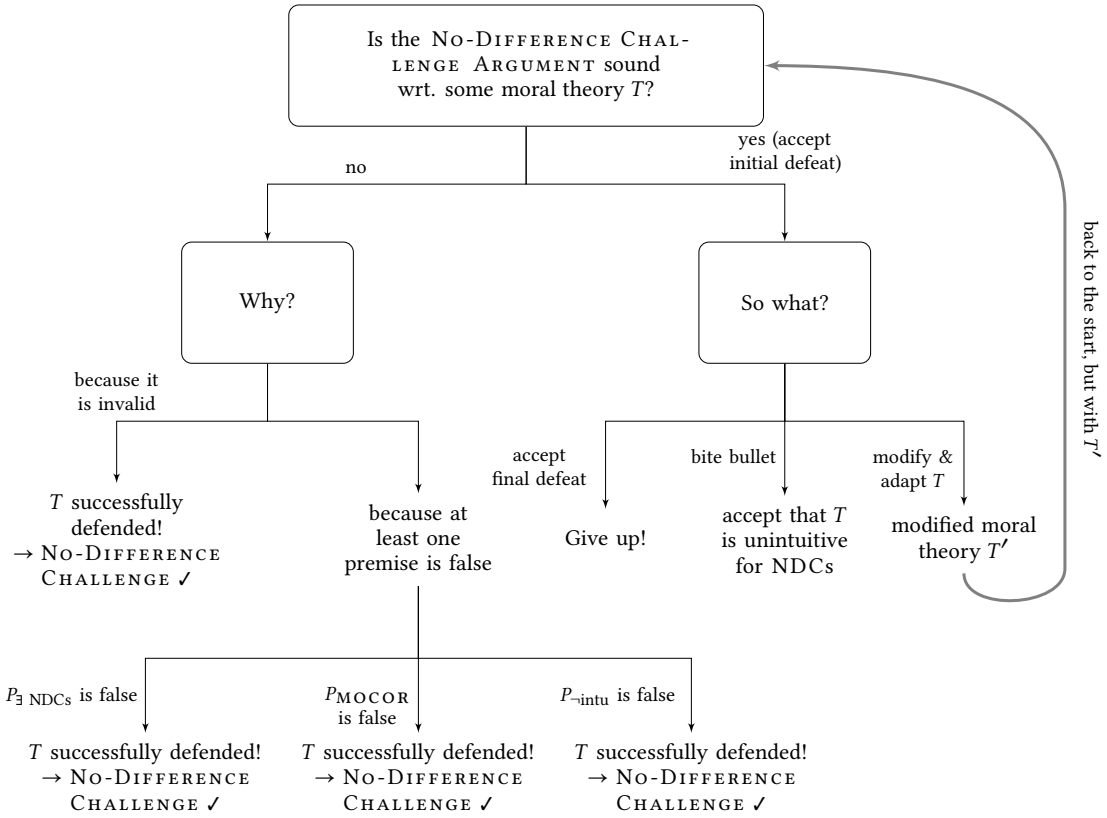


Figure 4.3: The solution space of the Challenge as No-Difference Challenge. Note that in the case of the No-Difference Challenge we may in principle allow for biting the bullet.

We have the same basic strategies available to handle the no No-Difference Challenge. Observe that bullet-biting might very well be an option in this case. It may seem painful, but by no means intolerable, to accept a theory's unintuitiveness (as $C_{\neg\text{intui}}$ states with respect to MOAC theories). Which theory does not have some counterintuitive implications, even systematic ones? Figure 4.3 shows the resulting solution space.

Finally, the Trilemma remains. Recall[93] that it arises from three apparently true propositions:

(H_1) Something wrong happens.

(H_2) If something wrong *happens*, then because someone *did* wrong.

(H_3) No one did wrong.

[93]The observant reader will have noticed that H_1 is formulated generically at this point, i.e., it is not applied explicitly to Two Factories.

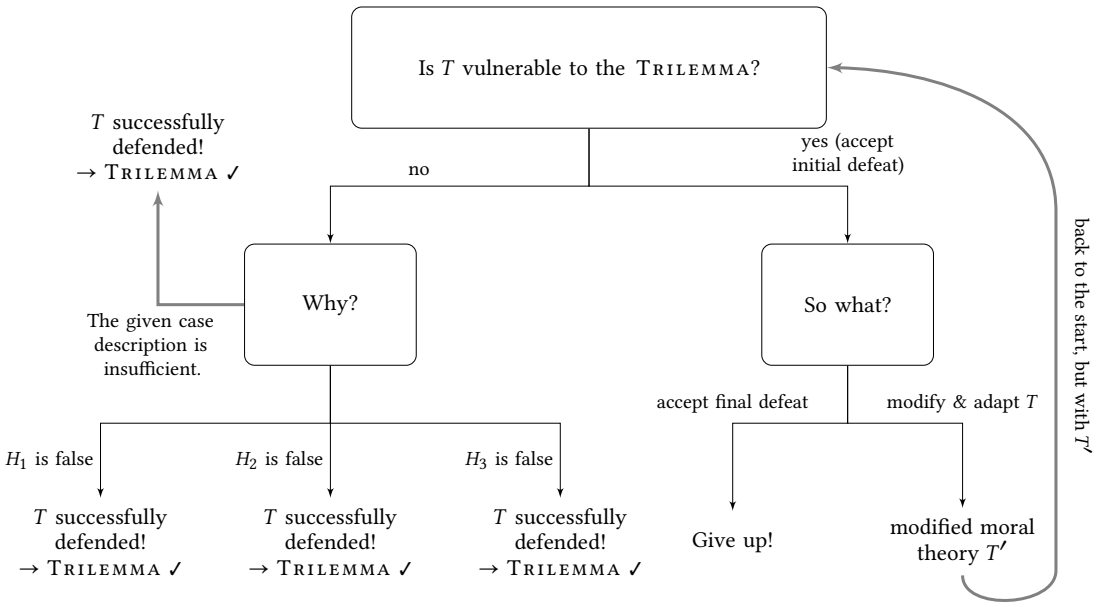


Figure 4.4: The solution space of the Trilemma.

Since not all of these three propositions can be true, any successful defense of some theory T against the Trilemma needs to explain why and how at least one of them is false relative to T. If this is not possible, the options remain to give up or modify T. Figure 4.4 shows the corresponding solution space for the Trilemma.

Finally, combining all three solution spaces produces the *overall solution space*, as shown in Figure 4.5. Note that any modification of the theory must start from the top, as modification that helps to overcome, for example, the Trilemma could 'reintroduce' vulnerability regarind the Argument.

With this structural groundwork in place, we now have a precise understanding of what it means for a theory to be *a* solution to the Challenge (and its subvariants). Next, we will address the question of what constitutes a *good* solution within the consequentialist aspirations of this thesis. However, before delving into this, I will first narrow the scope of this work using the framework of the solution space.

4.2 A Limitation: On the Exclusion of Cumulative Effects Cases from the Scope of this Project

In the introduction, I claimed that the Challenge arguably lies at the heart of some of the most pressing practical issues of our time, like the anthropogenic climate crisis. The central propositions in this context are

(22) It would be better to reduce carbon emissions significantly.

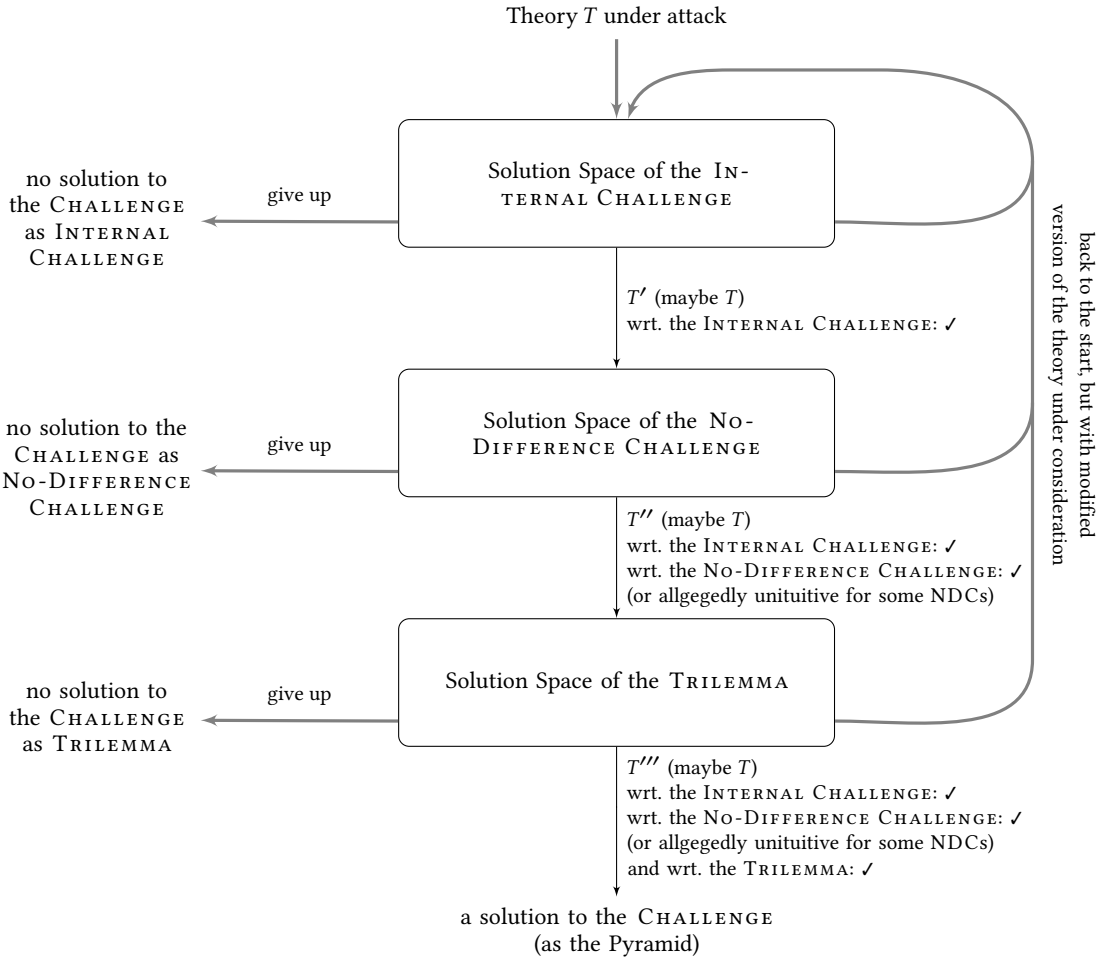


Figure 4.5: The solution space of the Challenge as the Pyramid.

(23) If sufficiently many people refrain from flying, become vegetarians, or switch to public transport, then this would cause a significant reduction in carbon emissions.

(24) An individual refraining from flying, becoming vegetarian, or switching to public transport reduces carbon emissions only morally insignificantly or even not at all.

Accordingly, in the case of the climate crisis, there seems to be a myriad of Cumulative Effects Cases as there are undoubtedly many overlapping and not just a single potentially troublesome combination. My individual contribution to the global disaster seems as negligible as it does now, even if my neighbor were to become a bike-loving vegan. But which combinations are troublesome

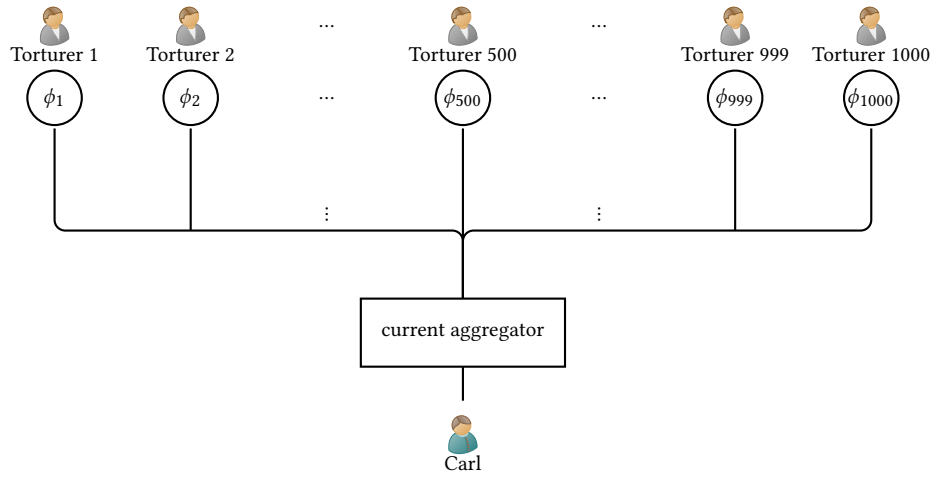


Figure 4.6: The setup of the Harmless Torturers case: 1000 torturers can each flip their switches s_1 to s_{1000}; the number of switches flipped determines the strength of the shock Carl receives.

exactly? And under what circumstances? Are there, for example, threshold values above which my switch from car to bicycle and from beef to tofu *would* make a difference?

Furthermore, real-life cases are especially challenging as they come with several *empirical* issues. How many CO_2 molecules make what difference in temperature where on earth for how long? How many of them make what kind of natural disaster how much more likely? This list could probably be continued almost indefinitely. Even if, at the very least, there are statistical correlations, they make the cases less than clear. For example, somewhat simplistically, one ton of CO_2 emitted is associated with losing three square meters of sea ice (Notz and Stroeve 2016). What does such a finding say about individual (in)effectiveness? Would it be appropriate to assign to each and every agent their 'fair share' even if we do not find a causal link?[94] A project like the present one, i.e., a project that aims at *theoretic* and *conceptual* challenges, is grateful if it does not have to lose itself in the subject area of the empirical sciences. I thus will not, and cannot, scrutinize real-life candidates for Troublemakers.

But even if we leave real-world cases aside and circumvent these empirical complications as far as possible, Cumulative Effects Cases still come with several conceptual difficulties. Recall

[94]This is, roughly, Glover's so-called "Share-of-the-Total View" (Glover and Scott-Taggart 1975). Parfit dismissed this view convincingly (Parfit 1984, chapter 3, section 25; and so we do not need to consider it in this chapter in any more detail, but see Figure 4.12 on page 128 for a classification of the approach.

Case 3.6 (Harmless Torturers) *Carl is wired to a torture machine with a thousand identical switches. When none of the switches are flipped, no current runs through the machine, so Carl is in no pain. If all thousand switches are flipped, then a considerable current runs through the machine, and Carl is in tremendous pain (but no permanent damage is done to his body). But the flipping of any given switch increases the current only by a tiny amount (well below the perceptually discriminable threshold for pain) so that Carl simply cannot tell whether one switch more or less has been flipped—regardless of how many other switches have already been flipped. Finally, imagine that a thousand different people each control a single switch and must decide whether to flip it or not. None of them cares about Carl or feels any remorse, but each of them enjoys flipping switches.*

Clearly, the ideal scenario would be for a sufficient number of torturers to refrain from flipping their switches—though defining "sufficient" is nearly impossible due to the inherent vagueness of cumulative effects. Thus, even though it is a carefully designed thought experiment, Harmless Torturers is by no means a simple and straightforward case. It rests on the vagueness of certain relevant predicates and the psychology of perception. Conceptually, the struggle starts already when we think about the fact that Harmless Torturers is described such that each and every particular flip

1. brings some joy to the torturer flipping the switch,
2. does not increase the harm at all or increases it only insignificantly (i.e., some imperceptible harm is inflicted, or whatever else characterizes insignificance here),
3. and yet the sum of these (non-)contributions somehow aggregates to more harm than 'the joys of flipping switches'.

These 'aggregative facts' certainly can make one scratch one's head. Shelly Kagan even claimed that such cases are *impossible* as, at some point—which might be hard or impossible to determine—there *must* be a "*perceptible* difference" (Kagan 2011, p. 134, fn 13). Here is his (slightly adapted) *reductio ad absurdum*:

> By hypothesis, when the person is in state 0, they are in no pain. If we ask them whether they are in pain, they will answer "no." In state 1,000, they are in excruciating pain. If we ask them whether they are in pain, they will answer "yes." Suppose then we consider state 1. Since this is adjacent to state 0, the difference between state 0 and state 1 must be imperceptible. Hence, if we ask someone in state 1 whether they are in pain, they must give the same answer as they gave when the same question is posed with regard to state 0, that is, they must answer "no." (If their answer in state 1 differed

> from their answer in state 0 this would presumably indicate a difference in their perception of the two states, contrary to hypothesis.) Now consider state 2, which is, of course, adjacent to state 1. Since, by hypothesis, the two adjacent states are imperceptibly different, the answer to the question "are you in pain?" must be the same. But the answer to this question with regard to state 1 is "no," hence the answer with regard to state 2 must be "no" as well.

First, note that Kagan considers Harmless Torturers as a Sequential Case, i.e., there are "adjacent" states that can be compared to each other. But this is only one of several implicit assumptions Kagan's argument rests on. For instance, it rests upon the assumption that the indiscriminability of two states implies that they are *equally* good (and not just, say, *on a par*). Furthermore, the latter part of his reasoning assumes that the overall harm is the sum of "trigger harms", meaning the individual harms caused by flipping each switch. A slightly generalized version of Kagan's argument goes like this:[95]

Conceptual Impossibility of Cumulative Effects Cases

P_1^{Imp}: If Cumulative Effects Cases are conceptually possible, then for any pair of adjacent states, the neighboring state is not morally worse than the preceding state.

P_2^{Imp}: If, for all pairs of adjacent states, the neighboring state is not morally worse than the preceding state, then the end state cannot be morally worse than the initial state.

P_3^{Imp}: If Cumulative Effects Cases are conceptually possible, then the end state is morally worse than the initial state.

C^{Imp}: Cumulative Effects Cases are conceptually impossible.

Whether this (obviously conceptually valid) argument goes through or not certainly depends on whether P_2^{Imp} is true (given that P_1^{Imp} and P_3^{Imp} are true by definition or description of Cumulative Effects Cases). Suppose Kagan's argument is sound (or, to be more exact, even if C^{Imp} were established on different grounds). In that case, consequentialists could rightfully ignore Cumulative Effects Cases with respect to their handling of the CHALLENGE—because this class of Troublemakers is simply empty then. So, what can we say about P_2^{Imp}?

[95] See Nefsky 2011 and, in particular, Spiekermann 2014 and Hedden 2020 for detailed reconstructions and critical analyses of Kagan's argument. The latter two specifically examine the idea that being morally on a par does not necessarily imply being good.

The answer is not as easy as Kagan wants us to believe. In fact, it depends on several axiological questions and on how we interpret specific phrases that are involved—i.e., it becomes at least in part a question of modeling (E. N. Dzhafarov and D. D. Dzhafarov 2010a,b; Hedden 2020; Spiekermann 2014). It thus seems possible to bend the axiological part of one's theory in such a way that Kagan's argument fails. However, the necessary adjustments, restrictions, and extensions, mainly related to vagueness, lead to an arguably otherwise unnecessarily complicated theory of value aggregation. Consequentialists might rightfully ask: Is it truly the role of consequentialism to set traps for itself? Should it not instead be the responsibility of those who attack consequentialism to precisely specify such an axiology and, moreover, to justify its plausibility?

In conclusion, considering Kagan's impossibility argument, it seems reasonable for the consequentialist to call for a shift in the burden of proof. Before we proceed to further defenses, it is incumbent on those asserting the Challenge to clarify precisely what the issues are. As far as I know, this has not yet been successfully done. Consequently, I exclude Cumulative Effects Cases from my discussion. In terms of the solution space of the Internal Challenge, this means that I simply deny the existence of an entire class of potential Troublemakers, namely that of Cumulative Effects Cases. Therefore, I reject $P_{\exists \text{TROUBLE}}$ of the Argument, I reject $P_{\exists \text{NDCs}}$ wrt. to the No-Difference Challenge, and for the Trilemma, I also dismiss the descriptions of the corresponding cases as insufficient or inconsistent. This guarantees that we have a trivial path to take through the overall solution space of the Challenge in terms of the Pyramid for Cumulative Effects Cases—we do not need to investigate in any further way.

4.3 Criteria for Good Solutions

With our understanding of the solution space established and a clear idea of what constitutes a solution, we're now ready to delve into the qualitative aspects of these solutions. We need a set of criteria to decide *what makes a solution good or bad*—or, at least, what makes one solution better or worse than another. Since, in our setting, solutions are theories, this quest boils down to determining what makes better or worse theories. But let's be clear: we are not embark-

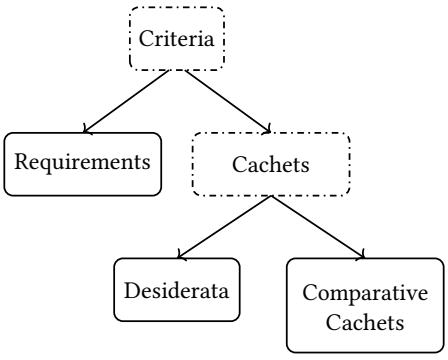


Figure 4.7: Schematic representation of the relation between the three different types of criteria. All criteria fall in the leaf categories of this tree.

ing here on an expedition to solve *all* problems of moral philosophy. Instead, our task here is more specific. We're interested in what makes one solution better than another in the context of this very project.

I introduce various criteria to evaluate and assess existing approaches to the CHALLENGE. There are three different classes of criteria: *requirements*, *desiderata*, and *comparative cachets*.[96] Figure 4.7 depicts the relationship between these types, and in the following, I sketch their characteristics and offer some examples. Along the way, I will make a selection of criteria that will be used for the assessment of solutions in the rest of this book. Mostly, I will mostly not provide detailed arguments in their favor but rely heavily on the intuitive convincingness of these criteria and on the fact that they have already been proposed elsewhere.

It should be noted that neither the classification nor the selection of the criteria is 'theoretically innocent'. Such innocence, however, would also not fit the project's overall aim. After all, the primary goal is to solve the CHALLENGE from the perspective of MOAC theories, and thus we are aiming here at a way to rank solutions based on their effectiveness in addressing the CHALLENGE from that particular perspective. Thus, our yardstick for quality is closely related to the overall success criteria of this project, i.e., developed with an eye on the formerly introduced solution spaces. As a result, the choice of the following criteria is heavily loaded with theoretical preconceptions.

4.3.1 Requirements:

Some criteria that I will call *requirements* are *necessary conditions of adequacy*. If a theory fails to meet any of these necessary conditions, it becomes disqualified and thus is removed from consideration as a plausible solution.

Some requirements are indisputable and convincing independently of further moral background assumptions.[97] Consider:

Property 4.1 (Deontic Consistency) *A theory is* deontically consistent *if and only if there is no decision situation such that, for some relevant context, some action is both right and not right according to that theory.*

A theory classifying the same action to be right *and* not right—not ambiguously, but as a substantial claim about its moral status in the very same sense—is inconsistent (cf. footnote 92). Similarly, consider:

[96] "Cachets", as I use the term, refer to qualitative markers or secondary criteria that help differentiate between candidate solutions when requirements and desiderata alone cannot resolve the comparison. They include considerations that, while not essential, may tip the balance in favor of one approach over another.

[97] Cf. Timmons 2001, p. 11.

Property 4.2 (Conceptual Deontic Consistency) *A theory is* conceptually deontically consistent *if and only if there is no decision situation such that, for some relevant context, some action is both right and wrong.*

That an action is both right *and* wrong is not a strict *logical* contradiction, though it arguably entails a *conceptual* one. After all, according to the *meaning* of the involved terms, a wrong action is necessarily one that is not right. Thus, a theory allowing for a situation where the same action is right and wrong would be one that is conceptually inconsistent.

Another hot candidate for a more theory-independent requirement is Methodological Individualism. Recall

Principle 2.1 (Methodological Individualism)
The primary bearers of deontic status are options of moral agents. Whatever has a deontic status and is not itself the option of a moral agent has this status merely in a derivative sense, that is, deontic status is a function of the deontic status of certain options of moral agents.

Other candidates for requirements *are* dependent on certain background assumptions. For instance, given an objective consequentialist perspective, one such requirement could be MH itself. Recall

Criterion 1.2 (Moral Harmony (MH) – tentative) *A moral theory is adequate only if it is true that if all agents act rightly (according to this theory), then they are guaranteed to produce the morally best outcome they could bring about together.*

However, since MH is an essential part of the CHALLENGE (as the INTERNAL CHALLENGE), we should not add it to our set of requirements. Either a proposed solution to the CHALLENGE (as the INTERNAL CHALLENGE) respects MH, or MH will be modified as part of it—and then this should be done for good reasons (depending on the path taken within the solution space sketched above, cf. Figure 4.2). To elevate it to the status of a necessary condition for a good solution would, therefore, unduly narrow the space of good solutions.

That said, there is another candidate for a requirement that depends on our moral philosophical background assumptions that we *should* add to the set of requirements. Since we seek objective consequentialist solutions, the theories we seek should qualify as such theories. In other words, they must fall under the definition at the end of Chapter 2:

Definition 2.3 (Objective Consequentialist Theory (formal)) *T is an* objective consequentialist theory *if and only if it embraces an axiological subtheory T_{Ax} with a valuation function* $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$ *and an objective consequen-*

tialist criterion of rightness T_{CoR} *such that, for all decision situations* $D \in \mathbb{I}$ *and for all* $\phi \in \Phi_D$ *:* $D, C \vDash_T \text{R}\phi$ *if and only if* $T_{\text{CoR}}(\phi)$.

A criterion of rightness T_{CoR} is objective consequentialist *if and only if, for all* $D \in \mathbb{I}$ *with* $\mathsf{D} \coloneqq \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O} \rangle$ *(with D's actual context C)* T_{CoR} *corresponds to a predicate* $\chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}$ *such that for all* $\phi \in \Phi$*:*

$$D, C \vDash_T \text{R}\phi \quad \textit{if and only if} \quad \chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}(\mathsf{Val}(\mathsf{Out}_C(\phi))).$$

Being an objective consequentialist theory is a *necessary condition* for any acceptable solutions (relative to the purpose of this project). After all, we came here to defend MOAC—and if a proposed solution to my CHALLENGE ceases to be a MOAC theory, then it is not what we are looking for. So, we include falling under the above definition in our list of requirements.

Note that requirements divide all moral theories $\mathcal{M}$ into two classes:[98] disqualified theories $\mathcal{M}_{\text{disq}}$ and candidate theories $\mathcal{M}_{\text{cand}}$, as

$$\mathcal{M} = \mathcal{M}_{\text{disq}} \sqcup \mathcal{M}_{\text{cand}}$$

with

$$\mathcal{M}_{\text{cand}} := \{\, T \in \mathcal{M} \mid T \text{ fulfills all requirements} \,\}.$$

4.3.2 Cachets

The second kind of criteria, called "cachets", may serve as tie-breakers for candidate theories $T \in \mathcal{M}_{\text{cand}}$. Cachets are of two kinds:

Desiderata: The criteria I refer to as "desiderata" are properties that a theory may or may not possess. Satisfying a desideratum makes a theory more favorable but does not imply its endorsement. Desiderata help to rank candidates and allow comparisons: Consider two candidate theories, T_1 and T_2. If T_1 satisfies an additional desideratum over T_2, then, other things being equal, T_1 is better than T_2. A plausible candidate for a desideratum is Regan's

Property 3.5 (Resolvability) *A moral theory is* resolvable *if and only if, for all decision situations and all relevant contexts, at least one option is right.*

Similarly, we might consider some of the principles implied by Resolvability, such as (cf. page 89):

Property 3.6 (Weak Deontic Completeness) *A moral theory is* weakly deontically complete *if and only if, for all decision situations and all relevant contexts, at least one action has a deontic status.*

[98] For those unfamiliar with this notation: $X = X_1 \sqcup X_2$ indicates X is the *disjoint union* of X_1 and X_2. More formally: $X = X_1 \sqcup X_2$ if and only if $X_1 \cup X_2 = X$ and $X_1 \cap X_2 = \emptyset$.

Another one is[99] (cf. page 89):

Property 3.7 (No Moral Dilemmas) *A moral theory is* free of moral dilemmas *if and only if, for all decision situations and all relevant contexts, not all actions are wrong.*

Alternatively, one might opt for the stronger version of the first one, i.e., (cf. footnote 83, page 89)

Property 3.8 (Deontic Completeness) *A moral theory is* deontically complete *if and only if, for all decision situations and all relevant contexts, all actions have a deontic status.*

Whether No Moral Dilemmas is considered a plausible candidate for an requirement or a desideratum certainly depends heavily on one's moral background assumptions. Deontologists may be able to live with the fact that they are committed to the existence of moral dilemmas, e.g., Kant, with respect to the murderer at the door (but see Cholbi 2009). However, from a consequentialist point of view, it is certainly a good candidate for a requirement (cf. Brown 2011).

It is not to be expected that a set of plausible desiderata will induce a nice and clean total order over candidate theories. Assume there are two sets of desiderata $\mathcal{D}_1$ and $\mathcal{D}_2$ which are at least partially distinct, i.e., with $\mathcal{D}_1 \setminus \mathcal{D}_2 \neq \emptyset \neq \mathcal{D}_2 \setminus \mathcal{D}_1$. What are we to do with cases where some candidate theory T_1 fulfills all the desiderata in $\mathcal{D}_1$ while some other candidate theory T_2 fulfills all the desiderata in $\mathcal{D}_2$? One could try to introduce some kind of hierarchy of desiderata, but I will not try to establish or defend such a hierarchy because I cannot see promising candidates for the required meta-criteria.

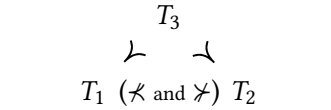


Figure 4.8: The partial ordering of the theories T_1 to T_3. We can only establish that T_3 is better than both T_1 and T_2. However, T_1 and T_2 are incommensurable.

Instead, let us think of desiderata as inducing a *partial* order $\prec \subseteq \mathcal{M}_{\text{cand}} \times \mathcal{M}_{\text{cand}}$ over the set of acceptable theories. Here is how this would work. Let T_1, T_2, and T_3 be three candidate theories, i.e., let all three fulfill all agreed requirements. Furthermore, let T_3 fulfill all desiderata in $\mathcal{D}_1 \cup \mathcal{D}_2$, while T_1 only fulfills desideratum $\mathcal{D}_1$ and T_2 only fulfills $\mathcal{D}_2$. The resulting partial order, including the incommensurability of T_1 and T_2, is captured in Figure 4.8.

[99] Note that, in principle, there could be theories that satisfy No Moral Dilemmas but not Resolvability. For instance, if, *pace* Consequentialist Standard View (cf. Section 2.2, page 21), some theory allows for situations where options have *no* (unconditional) moral status at all. (Note that the various forms of Deontic Completeness and the question of purely conditional moral status can be seen as applications of what Timmons calls "Determinacy", according to which a "moral theory should feature principles that, together with relevant factual information, entail determinate moral verdicts about the morality of actions, persons, and other objects of evaluation in a wide range of cases", Timmons 2001, p. 12.)

Comparative Cachets: Lastly, some 'criteria' are inherently comparative, relating theories rather than assessing them categorically. As an example, consider two traditional principles:

Principle 4.1 (Simplicity) *A theory T is, ceteris paribus, superior to another theory T' if T is simpler than T'.*

Principle 4.2 (Parsimony) *A theory T is, ceteris paribus, superior to another theory T' if T is more parsimonious than T'.*

Although neither principle plays a particularly prominent role in this project, a few comments are in order. First, it remains somewhat vague what exactly is meant by simplicity and parsimony. As an approximation, the following two suggestions may suffice for us. A theory is *simpler* than another if it requires fewer or, at any rate, less complicated propositions to explain an observation or phenomenon. (In the case of moral theories, the explanandum would be, e.g., the moral status of certain options or actions.) In contrast, one theory is more *parsimonious* than another if it requires fewer kinds of entities to provide such explanations. Sometimes parsimony is seen as an explication of simplicity, or it is assumed that being more parsimonious implies being simpler.[100] Certainly, introducing additional types of entities tends to make things more complicated in the first place. But in principle, nothing excludes that a theory provides simpler (less complicated) explanations through employing additional kinds of entities.

Both criteria are inherently comparative in that neither simplicity nor parsimony yields a useful unary predicate by itself. Even if we had an (intuitive) understanding of what makes a theory *simple* (or *parsimonious*), it seems hard to see to what extent this says anything about the quality of the theory as such. However, if one theory is *simpler* (or *more parsimonious*) than another, *ceteris paribus*, it is better.

Occasionally, it will be helpful to view extensional adequacy as a comparative cachet as well. Since, of course, we cannot know which actions are right in *all* decision situations—if we did, what purpose would moral theories serve?—we limit the evaluation to a *core* of 'clear, trivial, obvious' cases. (There is no need to define this set explicitly here; it can be specified later when concrete testbeds for particular theories are introduced.)

[100] For example, computer scientists occasionally direct me towards a passage from a textbook on machine learning in which we read (Gori, Betti, and Melacci 2023, p. 101, my italicization): "The parsimony principle (lex parsimoniae in Latin) is typically connected with classic Occam razor in philosophy, which states that entities should not be multiplied beyond necessity. Hence, whenever we have different explanations of the observed data, the *simplest* one is preferable."

Principle 4.3 ((Core) Extensional Adequacy)
Let

$$\mathfrak{I}_{\text{Core}} = \left\{ \left\langle D, C, \Phi_D^{\text{right}} \right\rangle \mid D \in \mathfrak{I}_T,\ \Phi_D^{\text{right}} \subseteq \Phi_D \right\}$$

be the core test set, *where C is the actual context of D and Φ_D^{right} is the set of actions assumed with certainty to be right in D. A theory T is,* ceteris paribus, *superior to another theory T' if T is extensionally more adequate than T' wrt. $\mathfrak{I}_{\text{Core}}$.*

This formulation of (Core) Extensional Adequacy leaves unspecified what *exactly* it means for some theory to be "extensionally more adequate" than another wrt. the core cases. Thus, in order to judge whether a theory T is more adequate than some theory T', I propose that we should just count the correctly identified right actions according to T (relative to $\mathfrak{I}_{\text{Core}}$), subtract the actions incorrectly assessed as right, and then do the same wrt. T' and compare the results. This formulation of (Core) Extensional Adequacy does not specify what it means *exactly* for a theory to be "extensionally more adequate" than another with respect to the core cases. To address this, I propose the following method: to determine whether a theory T is extensionally more adequate than another theory T', we count the actions that T correctly identifies as right (relative to $\mathfrak{I}_{\text{Core}}$), subtract the actions that T incorrectly assesses as right, and then perform the same calculation for T'. The comparison of these results will indicate which theory is extensionally more adequate.[101] This informal understanding will serve my purposes well enough.[102]

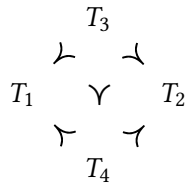


Figure 4.9: The partial ordering of the theories T_1 to T_4 is based on Simplicity alone. Regarding the desiderata, everything remains as in the previous example (cf. Figure 4.8); however, in terms of Simplicity, T_3 is simpler than T_4 (which is equally as simple as T_1 and T_2). Apart from Simplicity, T_3 and T_4 are otherwise equivalent, allowing us to rank T_3 above T_4.

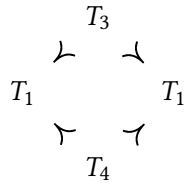


Figure 4.10: The partial ordering of the theories T_1 to T_4 is now based on Simplicity *and* Parsimony as criteria. As in the previous example, T_4 is more parsimonious than T_3 (which is equally as parsimonious as T_1 and T_2). Consequently, T_3 can no longer be ranked above T_4, nor can T_4 be ranked above T_3: they are incommensurable.

[101] For simplicity's sake, assume that all true (false) positives and true (false) negatives and all test cases $\mathfrak{I}_{\text{Core}}$ are equally important and thus count equally. However, see the observation in Footnote 62.

[102] For critical authors who demand a more explicit and formal notion, we can define:

Even if we assume that each comparative criterion might impose a total order over candidate theories, we still cannot reasonably hope that to be true with respect to the *entirety* of comparative criteria considered at once. A theory T might be simpler, yet less parsimonious, than another theory T'. As with desiderata, we thus may have to live with some sort of incommensurability.

For illustration, we can reuse our example from above and add another candidate theory, T_4. Let T_3 be simpler than T_4 (which is equally as simple as T_1 and T_2) while otherwise, T_3 and T_4 are on a par. Assume that there are no further cachets in the game. In this case, we can rank T_3 and T_4 both higher than T_1 and T_2 and T_3 higher than T_4 (cf. Figure 4.9). But now consider another comparative cachet, namely parsimony. Assume that everything is as before, but T_4 is more parsimonious than T_3 (which is equally as parsimonious as T_1 and T_2). Now T_3 has an advantage over T_4 and the other way around, and whether these advantages 'cancel out' might be undefined. This, then, means that we cannot say that T_3 is better than T_4, nor vice versa—nor that they are equally good (cf. Figure 4.10).

4.3.3 Some Notes on Criteria

All criteria—at least those applied in this book—fall into the three categories introduced above. As described earlier, requirements function as a 'filter' for theories, while the other two kinds of criteria help us to establish a partial order among acceptable theories. Thus, with regard to the question of whether one can find *the* best theory, a rather modest position seems advisable.

Furthermore, it is useful to think of requirements as being 'procedurally upstream' to cachets in the sense that, when considering a theory T, once we have found a requirement that T violates, we do not need to consider desiderata or comparative criteria. The latter two kinds, however, I will assume to be 'procedurally on a par', i.e., desiderata fulfillment does not, in principle, count more than having the upper hand over some theory with respect to some comparative criterion.

As already indicated in connection with the previously mentioned examples, there is a second, orthogonal dimension along which we can assess the criteria: their scope. Some criteria are so broad that they arguably are applicable to theories in general. Other criteria are criteria only for normative

$$\mathrm{EARel}(T,T') = \sum_{\left\langle D,C,\Phi_D^{\text{right}}\right\rangle \in \mathrm{I}_{\text{Core}}} \Big(\overbrace{\left|\mathsf{T}(D,C)\cap\Phi_D^{\text{right}}\right|}^{\text{'true positives' of } T} - \overbrace{\left|\mathsf{T}(D,C)\setminus\Phi_D^{\text{right}}\right|}^{\text{'false positives' of } T} - \Big(\overbrace{\left|\mathsf{T}'(D,C)\cap\Phi_D^{\text{right}}\right|}^{\text{'true positives' of } T'} - \overbrace{\left|\mathsf{T}'(D,C)\setminus\Phi_D^{\text{right}}\right|}^{\text{'false positives' of } T'}\Big)$$

In other words, then, some theory T is better than some other theory T' in terms of (Core) Extensional Adequacy if and only if $\mathrm{EARel}(T,T') > 0$; T' is better than T in terms of (Core) Extensional Adequacy if and only if $\mathrm{EARel}(T,T') < 0$; and the larger the absolute value, the stronger this advantage of one theory compared with another.

or even only for moral theories. Finally, there are criteria that are plausible criteria for consequentialist theories—and even some that are, if at all, plausible criteria for *specific* subvariants of it, like MOAC. Furthermore, some property might be a *desideratum* for theories in general, but a *requirement* for a specific theory or family of theories. We previously encountered this distinction in Subsubsection 3.5.2.2, where I argued that MOAC appears plausibly committed to MH, whereas MSAC is not. Figure 4.11 provides a Venn diagram illustrating an overview of the various types of theories and their relationships to one another.

That being said, I have not yet taken a final position on the question of what criteria to adopt—although the selection of the previous examples was not entirely arbitrary, of course. Accordingly, as Table 4.4 shows, I will start with those criteria that I consider to be quite plausible candidates (as I have also justified in passing above and in the previous chapter), especially from MOAC's point of view, of course. Nevertheless, in the following section, I will briefly explain why I think it would be *in*appropriate to elevate Deontic Completeness to the status of a requirement.

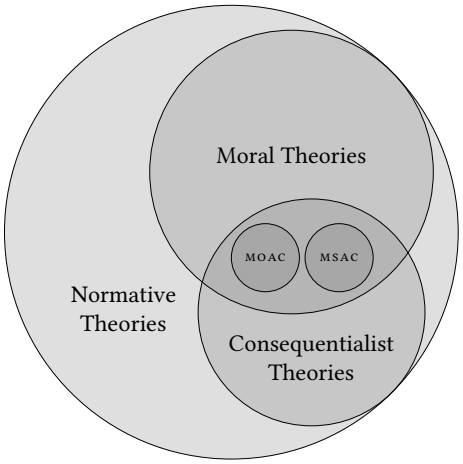


Figure 4.11: Rough-and-ready picture of the relations between normative, moral, and (different kinds of) consequentialist theories. (As an example of non-moral consequentialist theories, the reader might think of classical theories of instrumental rationality.)

4.3.4 A Counterexample Against Deontic Completeness?

The criteria listed in Table 4.4 were introduced with minimal supporting argumentation. Instead, I primarily relied on their intuitive appeal, given the consequentialist background assumptions. The potential pitfalls of this approach are highlighted by the following example,[103] which also illustrates why I included the properties No Moral Dilemmas and Deontic Completeness as desiderata rather than strict requirements. Consider the following case:

Case 4.1 (Evil Thorsten) *Thorsten finds himself with a spare Euro, a trivial amount he wouldn't even notice missing. A beggar approaches him, for whom that Euro would make a significant difference. However, Thorsten possesses a*

[103] The example is the result of a discussion with Thorsten Helfer and has been refined based on his feedback.

Requirements	Conceptual Deontic Consistency, Methodological Individualism, Being a MOAC Theory
Desiderata	Resolvability, No Moral Dilemmas, Weak Deontic Completeness, Deontic Completeness
Comparative Cachets	(Core) Extensional Adequacy, Parsimony, Simplicity

Table 4.4: A provisional list of criteria I will use below as yardsticks for assessing proposed solutions. Note that I add both the weak *and* the strong because I consider a theory that satisfies the strong version of some such property better: since every theory that satisfies the stronger version also satisfies the weaker one as a matter of logical or conceptual entailment, every theory that satisfies the stronger property fulfills two desiderata instead of only one.

particular aversion to acting as consequentialism prescribes. So much so that the personal disutility he experiences from adhering to consequentialist principles would be twice as impactful as the beggar's potential benefit.

This decision situation certainly seems somewhat suspicious, particularly due to the self-reference it potentially entails. Because once we evaluate the situation using a consequentialist principle, a self-reference appears: Thorsten's disutility is directly tied to his act of adhering to consequentialist reasoning. Is the liar paradox lurking around the corner (cf. Bolander 2024)?

At first, we might start our deliberation by constructing a naïve 'normal form'. Let us assume 10 as the utility the beggar gets from the Euro and, accordingly, a disvalue of -20 for Thorsten for doing what is right. Consequently, the sum of utilities for each case is as follows: giving the Euro when it is right yields a total of $+10 - 20 = -10$, while keeping it results in 0; if it is right to keep, giving yields $+10$, and keeping yields -20. Here is the corresponding table:

		Giving is right	Keeping is right
Thorsten	Give Euro	-10	10
	Keep Euro	0	-20

What would be the right thing to do for Thorsten if this were an adequate representation? We could apply the technique of conditionalization from Chapter 3. What would be the right thing for Thorsten to do if this were an adequate representation? We could apply the technique of conditionalization from Chapter 3. If it were right to give the Euro, then keeping it would result

in a higher overall utility (*Gesamtnutzen*). Thus, keeping would be the right thing to do, and giving would be wrong—a contradiction. Therefore, it cannot be that giving is the right thing to do. This reasoning might move us to claim that keeping must be the right thing to do. However, we then encounter a similar contradiction: If it were right to keep the Euro, then giving it would result in a higher overall utility. Thus, giving would be the right thing to do, and keeping would be wrong. Again, a contradiction. Consequently, it cannot be right to keep the Euro either.

The example might appear somewhat dubious, and some might argue that it is ill-posed. Admittedly, self-reference is always a red flag. However, it is important to note that this appears to be a well-defined decision situation: there is an agent with two options, and the consequences of these options are clearly specified. The issue arises not from the structure of the decision situation itself but rather from the features of the consequentialist framework. In this framework, the qualities of the consequences depend on what is deemed the right thing to do according to consequentialism—a theory that, in turn, determines the rightness of an action as a function of precisely these qualities. This creates an inherent regress, which is thus due to a feature of the theory rather than the decision situation. Consequently, the apparent 'impossibility' of properly describing the situation seems to stem from the consequentialist framework rather than from the case itself.

Consequentialists are left with two possibilities: either they argue that cases like Evil Thorsten are ill-posed but not due to any fault of their framework (and I do not see how they could), or they must choose one of the following three paths:

- Either they must live with the fact that at least one of the two options is right *and* wrong. In this case, they could argue that, for instance, giving is right and, therefore, keeping the Euro is better than giving it, such that giving it is wrong (and, hence, also not right). Doing so is to give up Conceptual Deontic Consistency (and, thus, also Consistency, at least if we stick to the (restricted) Consequentialist Standard View).

- Or they must accept that both options are wrong. Then the problem with Evil Thorsten vanishes, as the correct way to think about the case would be captured by

		giving and keeping are both wrong
Thorsten	Give Euro	10
	Keep Euro	0

This solution comes with both the pain of rejecting No Moral Dilemmas and the inconvenient question of how, then, it cannot be right to give the Euro. After all, this option has better consequences, so according to MOCOR, it must be the right thing—which, again, would violate Consistency.

- Or, finally, they must accept that the options have no moral status *simpliciter*. Then the problem with Evil Thorsten vanishes for similar reasons as above, as the correct way to think about the case would be captured by

		neither giving nor keeping is right
Thorsten	Give Euro	10
	Keep Euro	0

Again, they would need to explain why giving is not the right thing to do in light of MOCOR. But they could do so by excluding Evil Thorsten from the domain of MOAC theories, implying the surrender of Weak Completeness.

I am still unsure what lesson to learn from Evil Thorsten, but I think there are only two acceptable paths. Either camp MOAC finds a reasoned way to reject Evil Thorsten as being ill-posed, or they give up Weak Completeness. All other options come with unbearable theoretical costs. As argued above, no serious theory should allow inconsistent assessments, and no serious theory should deny its own criterion of rightness based on *ex ante* determinations. Because it's better to be safe than sorry, I suggest MOAC embrace Deontic Completeness as a desideratum rather than a requirement. For the remainder of my project, however, I exclude cases involving (implicit or entailed) self-references of moral theories.

Before presenting my own approach to the CHALLENGE in the next section of this thesis, I will gather evidence to support the claim that no convincing objective-consequentialist solution to the CHALLENGE has been proposed thus far.

4.4 An Unsatisfying Exploration of the Solution Space

I maintain that the currently proposed solutions to the CHALLENGE fail to provide a comprehensive resolution. However, I will not even *attempt* to substantiate this claim with a broad, overarching argument. This section neither aims to exhaustively cover all potential solutions nor to offer an in-depth analysis of any single approach. Such an undertaking would demand

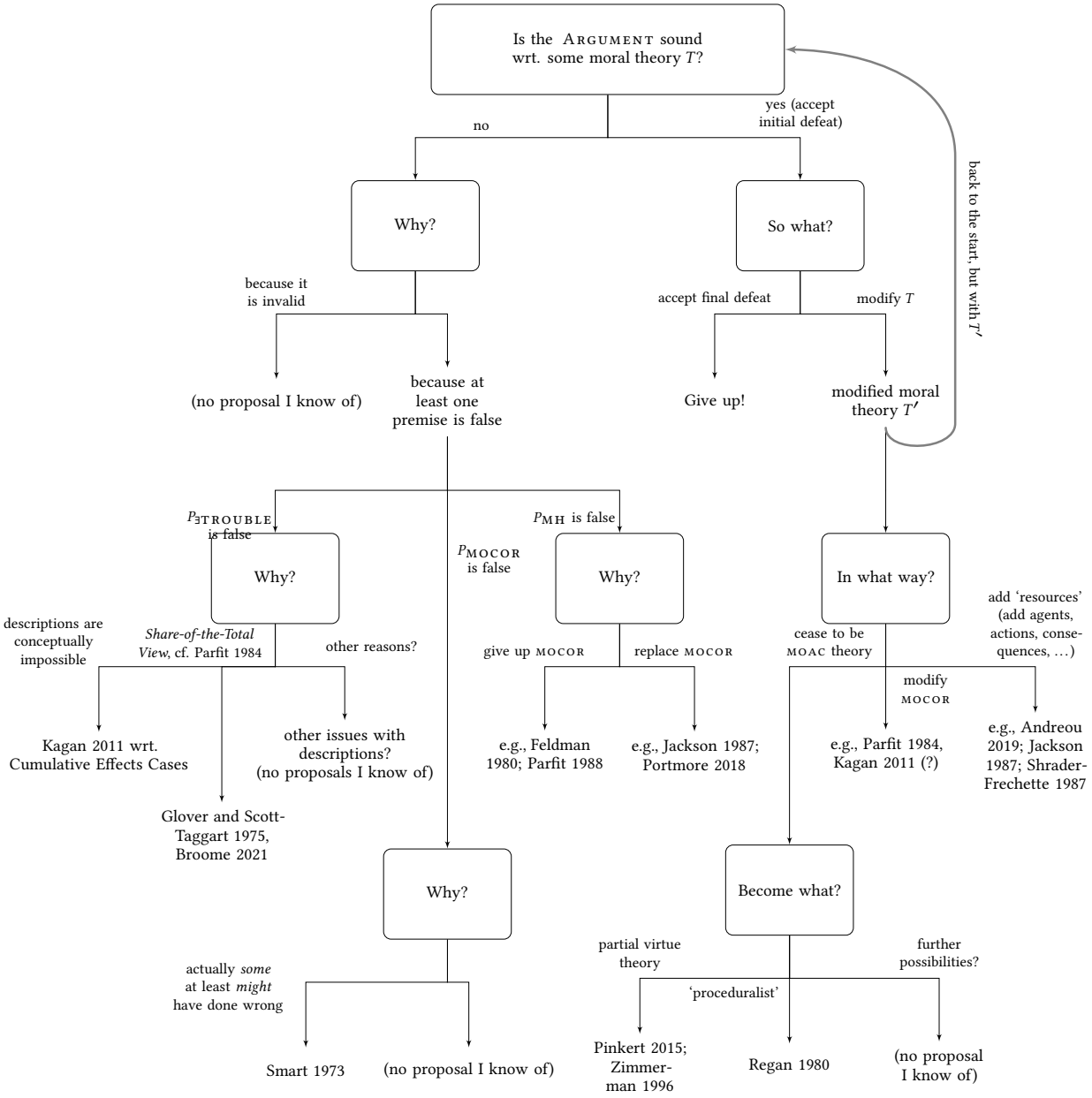


Figure 4.12: The solution space for the Internal Challenge (respectively, the Argument), annotated with a selection of (non-)solutions.

entire chapters, if not an entire book, and would unduly shift the focus of this work away from the development and presentation of my own approach.

However, Figure 4.12 organizes a carefully selected set of approaches to the Internal Challenge, the apex of the Pyramid. I have included approaches that are either well-regarded or potentially overlooked yet particularly illuminating. Notably, to the best of my knowledge, no one has critically examined the Argument by directly challenging its validity.

In the following, I highlight three carefully selected approaches. First, I address Shelly Kagan's 2011 proposal, credited with rekindling interest in the Challenge. However, Kagan's perspective can be set aside due to its early departure from core objective principles. Next, I briefly assess Felix Pinkert's 2015 contribution, which is particularly well-rooted in the overall debate but is ultimately dismissed for straying too far from fundamental act-consequentialist tenets. Finally, I explore an older and often overlooked approach by Frank Jackson (1987). Despite its limitations, Jackson's observation serves as an important bridge to the second part of this project where we will later, in addition, turn to Smart's 1973 approach).

Through this limited yet targeted exploration, I aim to illustrate that the Challenge remains unresolved, justifying the necessity of the subsequent segments of my thesis.

4.4.1 Kagan's Revived Discourse

Shelly Kagan's engagement with the Challenge has undeniably rekindled interest in, and discussion of, the Challenge. Even though Kagan's approach is emphatically systematic, it is not apparent which variant of the Challenge Kagan actually addresses, the Challenge as the Internal Challenge or the Challenge as the No-Difference Challenge. First, Kagan suggests that he considers moac when he says (Kagan 2011, p. 107) that "the consequentialist is indeed concerned solely with the production of the best possible results". We have already seen that this is not true for subjective variants of consequentialism (recall The Drug and msac). He then continues as follows (ibid., p. 107):

> There is [a] kind of case that might reasonably be thought to be problematic even from the perspective of consequentialism. These cases appear to have the following structure: A certain number of people—perhaps a large number of people—have the ability to perform an act of a given kind. And if a large enough group of people do perform the act in question then the results will be bad overall. However—and this is the crucial point—in the relevant cases it seems that it makes no difference to the outcome what any given *individual* does. And this is true regardless of whether others are doing the act or not. Thus, if enough people do perform the act the results are bad overall; but for all that, it remains true of each individual agent that it makes no difference to the overall results whether or not *they* perform the act in question.

Several details of this presentation can be criticized as slightly misleading: for one, Kagan writes of "overall bad" and not of suboptimal results (even overall bad results can please the consequentialist, if they are the best possible ones); or that Kagan here surprisingly speaks of acts "of a given kind",

although it does not matter what kind of action is performed, only what kind of dependency results in what kind of valuative profile. But these minor inaccuracies should not distract us from the fact that Kagan definitely has a version of the Challenge at hand. But which one?

I think that the most charitable reading is that Kagan is actually considering the No-Difference Challenge from a *subjective* point of view—even though this is not absolutely clear. First, consider a passage that apparently speaks against this reading (Kagan 2011, p. 107):[104]

> Intuitively, after all, in cases of the kind we are now turning to, the acts in question need to be condemned because of the results that eventuate from everyone's performing them.

This does not sound particularly subjectivist, though it might still be compatible with both a focus on the Internal Challenge and a focus on the No-Difference Challenge.

However, next to this, we find a passage that even suggests that Kagan is more interested in a version of the Internal Challenge (ibid., p. 108):

> The problem, in effect, is this: consequentialism condemns my act only when my act makes a difference. But in the kind of cases we are imagining, my act makes no difference, and so cannot be condemned by consequentialism—even though it remains true that when enough such acts are performed the results are bad. Thus consequentialism fails to condemn my act. In cases of this sort, therefore, consequentialism seems to fail even by its own lights.

Kagan's proposed *solution*, however, makes clear the subjective focus of his perspective (ibid., p. 119):

> How, then, can the consequentialist condemn my act? The key to the answer lies in the thought that it is only overwhelmingly likely that my act made no difference. It is unlikely, but possible, that it did make a difference—that my own act was the triggering act. But if it was, then of course it made a very significant difference indeed, for the triggering act brought about the various bad results. What we have, then, is a familiar case of decision making under uncertainty. I cannot know for sure that my act brought about the bad results—indeed, I can know that most likely it did not: but even when I discount the overall bad results for the high likelihood that my act did not bring them about, the net result of doing this remains negative. That is, my act has a negative expected utility. And that is why, from a consequentialist perspective, it should not be done.

[104] Again, it is misleading to frame the Challenge as being connected somehow to the fact that there is an option such that "*everyone*'s performing them". Sufficiently many agents suffice.

Kagan talks here about what one can *know* as an agent, what one has to *assume*, and that it is a decision under *uncertainty* that we encounter here. All these epistemic notions must be read as evidence that Kagan is not trying to make the methods of subjective consequentialism fruitful for MOAC, but is actually seeking to defend subjective consequentialism.

But then he cannot have the INTERNAL CHALLENGE in his crosshairs, for we have long since seen that this version of the CHALLENGE is not at all pertinent to subjective varieties of consequentialism—quite simply because they do not accept MH, that is, because they accept the compatibility of acting rightly and suboptimal outcomes, at least in cases involving, for instance, incomplete knowledge (recall Jackson's The Drug example again). Thus, for the sake of the coherence of his work, Kagan must be considered as being concerned with the CHALLENGE as the NO-DIFFERENCE CHALLENGE.

This is not yet to say, of course, that we cannot use Kagan's approach for MOAC as well. However, a moment's reflection reveals that it cannot be done *trivially*: Where would the probabilities required to compute expectation values come from if they are not subjective? Would they be objective probabilities? How would these fit into an objective consequentialist framework, and what commitments would their inclusion entail? Kagan offers no answers to such questions, simply because he apparently does not aim to defend objective consequentialism in the first place. Since that is the focus of my project, we can set this part of Kagan's approach aside for now.[105]

4.4.2 Pinkert's Modal Virtue Consequentialism

Unlike Kagan's account of the problem, we have already discussed Pinkert's account in some detail (Subsubsection 3.5.2.2). There can be no question about which version of the CHALLENGE he has in mind. Pinkert has even proposed his own explication of PMH, called On-the-hook (cf. Subsubsection 3.5.2.2), which is meant to go beyond MOAC in its applicability. In this respect, it is clear that Pinkert wants to solve the CHALLENGE as INTERNAL CHALLENGE. He makes his solution exemplarily explicit in the form of a new criterion of rightness[106] (Pinkert 2015, p. 982):

Principle 4.4 (Modally Robust Act Consequentialism (MRAC))
An agent acts rightly if and only if the agent acts optimally in the actual world,

[105] My interpretation aligns with Brian Hedden's 2020, who defended and expanded upon Kagan's approach. Additionally, we will revisit the question of how far expected utility–based considerations can be applied in the context of the CHALLENGE in the next part of this project.

[106] More precisely, Felix Pinkert formulates a 'criterion of oughtness', which differs slightly from the formulation presented here. His version states: "An agent ought to act optimally in the actual world, [...]".

and it should be such that for all possible combinations of the actions of other agents, if that combination were instantiated, they would act optimally in these circumstances.

MRAC is an intriguing blend of act consequentialism with modal considerations (dealing with possibilities across different conceivable worlds). The first part, according to which an action is right if and only if, in the actual world, it leads to the best possible outcome, is basically MOCOR. The second part, which concerns 'modal robustness', introduces a stringent requirement: not only should the agent's action be optimal in the real world, but it should also be such that in every conceivable scenario where other agents might act differently, the agent would still act optimally under these circumstances. This ensures that the agent's action isn't just accidentally right but is robustly optimal across various possible contexts.

In my opinion, Pinkert's solution has one crucial flaw: although at first glance the condition *looks* like an objective-consequentialist one, on closer inspection, it is not. Recall the formally precise condition for objective-consequentialist theories:

Definition 2.3 (Objective Consequentialist Theory (formal)) *T is an* objective consequentialist theory *if and only if it embraces an axiological subtheory T_{Ax} with a valuation function $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$ and an objective consequentialist criterion of rightness T_{CoR} such that, for all decision situations $D \in \mathbb{I}$ and for all $\phi \in \Phi_D$: $D, C \vDash_T \text{R}\phi$ if and only if $T_{\text{CoR}}(\phi)$.*

A criterion of rightness T_{CoR} is objective consequentialist *if and only if, for all $D \in \mathbb{I}$ with $\mathsf{D} := \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O} \rangle$ (with D's actual context C) T_{CoR} corresponds to a predicate $\chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}$ such that for all $\phi \in \Phi$:*

$$D, C \vDash_T \text{R}\phi \quad \textit{if and only if} \quad \chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}(\mathsf{Val}(\mathsf{Out}_C(\phi))).$$

Thus, for a criterion of rightness to be genuinely objective consequentialist, it needs to be based *solely* on the valuative profile of the situation (i.e., $\mathsf{Val}(\mathcal{O}_{D,C})$) and the value of the action under consideration (i.e., $\mathsf{Val}(\mathsf{Out}_C(\phi))$).

There cannot be such a predicate for Pinkert's MRAC. The modal robustness component makes it so that this predicate would also need to account for the optimality of actions *across* possible worlds. Assessing whether an agent acted rightly thus requires more than looking at the value of that action and the quality of the possible outcomes of the decision situation given the *actual* context. It also needs information about what action *would* have been taken instead if the other agents had acted differently. Therefore, it needs a reference to a disposition (or something similar) of the acting agent, and hence, it needs another parameter, namely the agent of the action. This creates a kind of virtue-theoretical character (in the sense of the distinction in Section 2.3), and

Pinkert's theory thus ceases to be a purely objective-consequentialist one—and therefore, in the sense of this project, is not a viable solution (Figure 4.12). Simply because (in Regan's *lingo*), MRAC is not 'exclusively act-oriented'.

Pinkert himself is well aware of this connection between the modal part of MRAC and virtuous character traits (as he tells us already in his abstract: "I interpret this Modally Robust Act Consequentialism as Act Con- sequentialism plus a requirement of moral virtue", Pinkert 2015, p. 971). My aim here is merely to clarify the issue and explicitly identify Pinkert's solution as incompatible with foundational principles of pure objective consequentialism—and, by extension, with the goals of this project. I don't think Pinkert himself would have much of a problem with this. And even I would agree that, as long as there is no *purely* objective-consequentialist solution to the CHALLENGE, Pinkert's approach is probably the most acceptable one for proponents of MOAC—even if it would imply making a move within the theoretical landscape. However, I promise to present, in the second part, a solution that makes such a move completely unnecessary. At this point, it should simply be noted that accepting Pinkert's solution would require a significant departure from foundational principles of consequentialism.

4.4.3 Jackson's Collectivism

Finally, we turn to Frank Jackson's take on the CHALLENGE. Jackson's position is essentially a direct response to Parfit's proposal in *Reasons and Persons* (Parfit 1984),[107] where Parfit suggested the following extension of the consequentialist framework (ibid., p. 70):

> (C7) Even if an act harms no one, this act may be wrong because it is one of a *set* of acts that *together* harm other people. Similarly, even if some act benefits no one, it can be what someone ought to do, because it is one of a set of acts that together benefit other people.

Jackson believes this to be mistaken and brings up several quite convincing examples against (C7), which need not interest us here in detail. Ultimately, Jackson suggests (Jackson 1987, pp. 100–101) that the CHALLENGE is rooted in our "tunnel vision", i.e., that we are "restricting ourselves, without fully realizing it, to the individual actions" while, instead, we should "enlarge the class of actions which may be morally evaluated to include group actions as well as individual actions" because then "we can say that the agents' group actions, though not their individual actions, are wrong." In sum, Jackson's position is this (ibid., p. 101):

[107] Parfit's reaction, in turn, is directed primarily at Glover (1975) and even more so at Regan (1980). To Jackson, in turn, Parfit has reacted with an unpublished piece, also with an approach that goes more in the direction of 'biting the bullet' (Parfit 1988). See also the overview in Figure 4.12.

> Parfit wants to enlarge our conception of what makes an action wrong with his guilt by association theory. I am suggesting that we respond to the difficult cases for the Difference Principle by enlarging our conception of what kinds of actions can be wrong (and right).

What may *prima facie* look like an elegant solution, comes with several weaknesses. First, Jackson's approach blatantly violates a requirement:

Principle 2.1 (Methodological Individualism)
The primary bearers of deontic status are options of moral agents. Whatever has a deontic status and is not itself the option of a moral agent has this status merely in a derivative sense, that is, deontic status is a function of the deontic status of certain options of moral agents.

This theoretically well-founded principle seemed *prima facie* justified and a valuable building block in quite a lot of theoretical work: no action without agent—and, maybe even more importantly, no wrong-doing (or right-doing) without agent as well. Jackson *at least* owes us a good argument as to how a departure is justified and not just an *ad-hoc* dodge.

Second, it is not clear how Jackson's approach could help with the Challenge as the pmh-based Internal Challenge. Since pmh is based on the idea that consistently right action must be accompanied by the best results, the question must be asked how a wrong combination of actions to which no agent corresponds should help. In the end, even if we found a wrong *combination*, we would not necessarily have found a wrong *action*.

However, Jackson's approach might still provide a pathway to resolving the Challenge, both as the No-Difference Challenge and, more specifically, as the Trilemma, since H_2 would, in that case, simply be false. Recall:

(H_2) If something wrong *happens*, then because someone *did* wrong.

According to Jackson, any pre-theoretic acceptance of H_2 would just be explained by our 'tunnel view'. This seems not implausible to me. Still, it does not help with the Challenge as the Internal Challenge.

Third, the question arises of how the moral status of a combination of acts relates to the moral statuses of the individual acts. (Depending on Jackson's answer, this could potentially help to address the first issue as well.) Jackson himself had much to say on this topic, and in light of the second part of this book, it is worthwhile to briefly examine his perspective.

Before I let his thoughts blossom, however, it should be summarized first that while Jackson's approach seems simple and charming, it comes with a lot of theoretical baggage. In a sense, similar things apply to Jackson's approach as to Pinkert's: The approach is not yet definitely out of the game. If there is

no better solution, the work of spelling it out in detail might be worthwhile for proponents of MOAC. But the cost would be high, and the necessary justificatory work seems quite extensive to me. Furthermore, it is likely that *new* challenges and counterexamples, arising from the new theoretical liabilities, are going to be found. So, overall, it should be worth it to look for an approach that comes with fewer commitments.

But before we use this impulse to move to the second part of my project, let us return briefly to the connection between the moral status of a combination of actions and the moral statuses of these individual actions: Jackson claims (1987, p. 101) that "the moral standing of a group act can be partially or totally at variance with the standings of its constituents". Here is an example Jackson uses to make his point (ibid., p. 102):

Case 4.2 (Intersection) *You and I approach an intersection from different directions. I have the right of way, so that what ought to happen is that you give way together with my driving straight on.*

The following normal form may serve as an appropriate representation of Intersection as envisioned by Jackson:

		You	
		drive	stop
I	drive	worst	best
	stop	second-best	second-worst

Intersection is indeed an asymmetric Coordination Case.[108] Much more interesting is how Jackson *argues* for what is right and wrong in this case, given a certain combination of actions, and what he thinks should be said about the moral status of that combination. Here is Jackson arguing that, given that we both drive, my driving is wrong, but the combination of [you stopping and me driving] would be right (ibid., p. 102):

> What in fact happens is that you do not give way, and would not regardless of what I do; so that were I to drive straight on, there would be an accident. What ought I to do? Drive straight on, consoling myself with the thought that I will be able to say from my hospital bed that I was in the right? Obviously, what I ought to do is stop. The position, then, is that the right group action

[108] The property of (a)symmetry will be discussed in greater detail later. For now, an intuitive understanding will suffice: a case is symmetric if exchanging the agents does not alter the moral qualities of the consequences associated with their options. This applies to Two Factories but not to Intersection. In Jackson's construction, it is better for the agent with the right of way to drive while the other stops, rather than the reverse. For instance, if *you*, rather than *me*, had the right of way, it would be better for you to drive and for me to stop, effectively reversing the roles in the original example.

> is your stopping together with my driving on; the right action for you is to give way, and the right action for me is to stop. But if the right action for me is to stop, the wrong action for me is to drive on. Hence, we have a group action—your stopping together with my driving on—which is right, which nevertheless has a constituent action—my driving on—which is wrong.

It is to be anticipated that some readers will object to Jackson's argument along the following lines: "Of course, it's right for me to stop, but that's *because* you're not doing what you're supposed to do—i.e., you're not doing what is right for you. And yes, in a certain sense, it would be right for you to stop and for me to drive (because I have the right of way or for some other reason), but this *presupposes* that this combination is still an option at all (for whom, exactly, by the way?). However, once it is determined that you drive on, that combination is no longer on the table." This kind of objection, I believe, underscores the necessity of distinguishing carefully between the *initial state*, where it is not yet determined what anyone will do, and the state once *some action has been decided, initiated, or completed.* This distinction underline the need for *dynamics* in the analysis—a conceptual space that consequentialists currently lack in their understanding of Coordination Cases. Therefore, since camp MOAC does not yet possess this conceptual resource, Jackson can effectively counter this anticipated objection with the following straightforward move (Jackson 1987, p. 102):

> I have been surprised by how often I have met the following response to this sort of example. "The argument turns crucially on the claim that I ought to stop. But I ought to stop *only because* you do not do as you ought, namely, give way, and so all that is really true is that I ought to stop given you do not give way." However, the sketched alternative position is inconsistent: "P only because Q" entails P! To grant that it is true that I ought to stop is true only because you do not give way is *ipso facto* to grant that it is true that I ought to stop.

Of course, Jackson is correct that "p because q" implies that q.[109] That said, this observation does little to substantiate his claims. To fully grasp why Jackson's argument falls short, however, we need a more robust framework for Coordination Cases and collective decisions—one that enables reasoning about sequences of actions, facilitates the decomposition of Troublemakers into sets of individual decision situations, and provides the conceptual space to assess combinations of actions. Developing such richer understanding sets us on a path that offers unexpected insights, including fresh perspectives on the CHALLENGE and its validity, while paving the way for a more sustainable resolution to the challenges of consequentialism in multi-agent settings. This journey begins in the second part of this thesis.

[109] For a well-developed logic of *because*, see, e.g., Schnieder 2011.

Part II

The Real Challenge (and How to Solve It)

Overview of Part 2

In the second part of this book, I develop my very own approach to the Challenge. I will do so in two steps. The first step reveals a serious misconception underlying the Challenge as reconstructed in the first part of my project. I suggest understanding the Challenge rather as a symptom of a deeper cause, i.e., as a result of a move consequentialists have tacitly made in order to avoid *another* challenge I will call the Real Challenge. The second step, then, is to offer a solution to the Real Challenge that does *not* result in the Challenge. Proceeding in this way, I master several challenges for consequentialism at once: I demystify and deconstruct the well-known and much-discussed Challenge and, at the same time, solve the deeper-lying and so far only sporadically discussed Real Challenge. In the end, both challenges are off the table and Multi-Agent Consequentialism is born.

Building on the formalism introduced earlier, I extend it to the domain of collective decision situations in Chapter 5. In Chapter 6, I demonstrate that something is fundamentally flawed with the Challenge by exposing an equivocation in the Argument, ultimately arguing that it should be deemed invalid. Roughly speaking, it is questionable whether the right actions derived in P_{MOCOR} really stand in conflict with the missing wrong action required in P_{MH}. A closer look at the two premises raises the question of whether the same context of evaluation is relevant within the two premises. P_{MOCOR} explicitly assesses retrospectively, i.e., *ex post*: given what the others did (or, alternatively, by presupposing what the others will do, …; cf. Subsubsection 3.5.2.3), none of the agents could have changed anything for the better. For this reason, each of them has acted rightly, given the actions of the other agents, but only in retrospect. P_{MH}, however, apparently results from an aspiration to 'guide' (future) behavior, namely, in the direction of the best outcome. I thus argue that the best reconstruction of the Argument is invalid.

I close the chapter with the somewhat surprising conclusion that this result is not as good for MOAC advocates as one might initially think. After all, this result is a Pyrrhic victory, revealing what I call the Real Challenge. This challenge consists, roughly, in the diagnosis that MOAC does not give *incorrect* assessments but rather no (unconditional) assessments at all. I argue

that the Real Challenge is even more severe than the original Challenge, as it implies that: (i) MOAC is critically deontically incomplete, and (ii) MOAC theories ultimately fail to uphold the spirit of PMH: if there are no genuinely right actions, we cannot hope that morality 'highlights' or points toward the way to the best possible outcome. Revisiting Regan's 'impossibility result', I conclude the chapter by explaining why we should understand the Challenge as a symptom of the consequentialists' approach to filling the deontic gaps in certain collective decision situations. In the remainder of this part, what remains is to find a way for camp MOAC to fill the gaps *without* invoking the Challenge.

In Chapter 7, I propose that consequentialists may have overlooked—or perhaps forgotten—a plausible candidate for what might be termed *intermediate outcomes* in collective decision situations: the decision situations of other agents. Adopting this Intermediate Outcomes Approach (or simply "Approach" for short) introduces brand-new material to the consequentialist workbench. I demonstrate that this fresh perspective enables a unified understanding and representation of Coordination Cases and Sequential Cases through what I term *general extensive forms*. This approach brings new clarity to camp MOAC, which now faces the task of deciding *how* to fill the deontic gaps and thereby overcome the Real Challenge—ideally without falling back into the Challenge. To aid in this effort, I introduce what I call *collective amendments*: mathematical, formal methods designed to morally valuate the newly identified intermediate outcomes. These amendments, thus, serve as the new tools for the consequentialist workbench, complementing the new material already uncovered. However, without a clear framework for assessing these tools, we are not yet in a position to determine which amendments should be adopted.

To overcome this last remaining hurdle, I revisit the PMH in Chapter 8. Based on yet another kind of Coordination Case, I argue that the current explication, MH, must be relaxed because a theory adhering to MH would necessarily conflict with an even more basal and persuasive principle, namely, the principle of Normative Supervenience, a principle logically entailed by the very definition of objective consequentialism. This opens up a new perspective that allows us to explicate the overarching goal consequentialists should aim for when filling the formerly diagnosed deontic gaps. Building on the concept of calculating the expected value of adopting specific amendments, I propose a formal framework for ranking amendments, inspired by the notion of policies in formal decision theory. After defining a 'testbed' of key structural types of collective decision situations, I perform calculations to identify a preferred amendment—and to determine a winner.

Finally, I conclude my project in Chapter 9 by highlighting potential avenues for future research. I close by challenging deontologists, who now

seem to be the ones grappling with collective contexts, to consider moving closer to—at least a bit—camp MOAC to benefit from the novel conceptual possibilities introduced in this work.

As in the first part, it is prudent to address some formalities upfront to avoid introducing new concepts piecemeal throughout the argument, as this could disrupt the flow. The next chapter, therefore, focuses on extending decision situations to collective contexts. These at least semi-formal definitions and concepts will play a more prominent role in this second part than formalisms did in the first.

Chapter 5

Preliminaries II

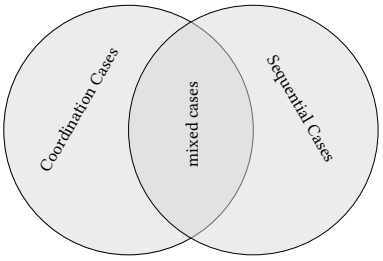


Figure 5.1: Kinds of collective decision situations.

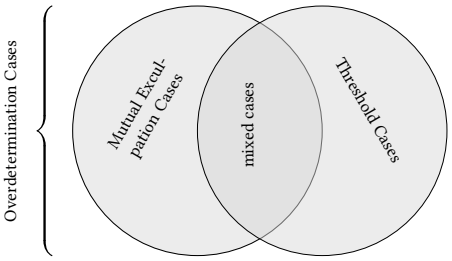


Figure 5.2: The set of all Troublemakers as tackled by my project (see Section 4.2). As explained earlier, it suffices to focus on pure cases. Recall that Mutual Exculpation Cases are Coordination Cases while Threshold Cases are Sequential Cases.

At this point, we have identified two types of collective decision situations: Sequential Cases and Coordination Cases (cf. Figure 5.1 and Figure 5.2 for their respective Troublemakers), along with numerous specific instances of these categories. So far, quite in line with the debate on the CHALLENGE, I have assumed that we can simply apply MOAC to the agents' decisions in such collective situations, a view I shall call Compositionalism. As already marked earlier, behind this practice is the implicit assumption that, in a sense that deserves to be dragged to light, collective decision situations can somehow be separated into individual decision situations (or, at least, that one can identify individual decision situations for each agent).

On closer inspection, however, the situation is less straightforward than it might initially appear. In collective decision situations, we are not necessarily guaranteed to identify 'proper', i.e., unconditional, consequences for individual agents. This raises the question of whether Compositionalism holds true, or whether we should instead adopt the Genuine Kind View, which asserts that at least some collective decision situations are of an entirely distinct nature. If the latter is correct, it becomes unclear what MOAC has to say about such cases. This uncertainty, therefore, proves to be a pivotal issue in addressing the CHALLENGE. Consequently, the second part of this work begins with a thorough examination of the relationship between collective and individual decision situations.

To gain a clearer understanding of this question and the arguments encountered so far, I will introduce a number of concepts and (semi-)formal notions. These tools will help me to concisely explicate key properties of collective decisions, particularly the Triad of *maximality*, *order invariance*, and *symmetry*. First, these properties will allow me to precisely narrow the focus of my project to a manageable scope. Second, they will serve as the foundation for a systematic treatment of the issue later on.

Finally, I will examine what MOAC (and, more specifically, MOCOR) can truly offer in addressing Coordination Cases, along the way introducing the concept of the *decomposability* of collective decision situations.

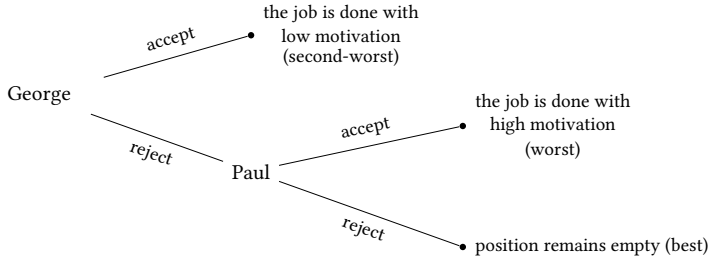
5.1 Collective Decision Situations

While normative and moral questions primarily concern individual decision situations, the CHALLENGE is raised by *collective* decision situations. Up to now, I have pretended (following the existing discourse) that trivially, each of the agents involved in a collective decision situation faces his own individual decision situation. We will see that this is not at all an innocent assumption.

We have already encountered several collective decision situations in this book, and we will meet several more. Thus, we have already developed an intuitive understanding of collective decision situations that we can now refine toward a more precise definition. Recall the rough-and-ready definition from Chapter 3:

Definition 3.1 (Collective Decision Situation – tentative)
A collective decision situation *is a situation in which multiple agents are each presented with multiple options, and within a given context, each combination of actions has an associated consequence.*

In light of the situations we have encountered thus far, this definition needs some slight modifications. Most notably, we should make conceptual room for the possibility that, in certain cases, agents face their decision situations only if other agents have already decided or acted in specific ways. Consider, for example, Job Market, which was represented in extensive form as follows:



It seems appropriate to say that in this Sequential Cases, there are two agents, George and Paul, but only George is *guaranteed* to be put into a situation where he has to decide. Paul may have no choice to make at all because if George takes the job, it won't be offered to Paul. Thus, the combination of George accepting the job and Paul likewise accepting it is not a possible combination of actions according to Job Market. However, both agents are needed for a sufficiently complete description of the situation with respect to the possible, relevant consequences. For example, both the action combination of George refusing the job and Paul accepting it and the 'combination' consisting of merely one action, namely George accepting the job, are associated with genuine outcomes.

Let it be agreed that a combination of actions is called *proper* (relative to a concrete collective decision situation) if there is a specified outcome for it. Further, I will describe choices like Paul's such that an agent is *potentially* presented with multiple options. We can then define collective decision situations a bit more precisely like this:

Definition 5.1 (Collective Decision Situation) *A* collective decision situation *is a situation in which multiple agents are each (potentially) presented with multiple options, and there exist some* proper *combinations of actions within a given context—combinations of actions that have associated consequences.*

None of this excludes the possibility of hypothesizing about purely hypothetical actions—those an agent might have taken if given the choice. Hypothetical actions can, in fact, play a significant role in moral assessments and related considerations. As we will see, understanding the interplay between potential choices, proper combinations, and the moral assessment of, and based on, hypothetical actions is essential for consequentialists to address the Challenge. However, achieving such an understanding is far from straightforward. In the remainder of this section, I will lay the groundwork that will serve us throughout the rest of this part in developing that understanding.

A note at the outset: For the sake of readability and clarity in many of the following considerations, I will not attempt to address the Challenge in its most general form. Instead, I will apply certain simplifications and limitations. Even after thoroughly reviewing my results, I am confident that generalizing them would involve little more than technical adjustments and formal exercises. Ultimately, this is a dissertation in analytic philosophy, one that is unafraid to engage with formalisms. However, it is not intended to be a formal treatise in mathematical multi-agent ethics.

The first of these limitations has already been mentioned earlier. I primarily focus on the 'smallest' collective decision situations—those involving two agents and two options each. This simplifies representation and yet everything discussed here can be generalized to cases with any number of agents

and options. Such generalization is achievable through relatively straightforward inductive reasoning, provided we establish a sufficiently suitable and robust formal structure. I will elaborate on this point when I summarize my approach at the end of this book. Importantly, this generalization requires a formal framework capable of supporting mathematical operations accordingly.

Before I turn to introducing such structure through formal specifications and shorthands similar to those introduced earlier in the context of individual situations, we should pause for a moment and ask whether we *really* need to introduce another kind of decision situation. In other words, aren't individual decision situations sufficient, possibly just complemented with some structural information about dependencies between agents, actions, and consequences?

Two questions arise here, only one of which can be immediately settled. First, can there be additional types of decision situations beyond individual and collective ones? In other words, are there possibly even more *kinds* of decision situations? This question, I believe, can be answered in the negative, as no compelling candidates come to mind, nor have I encountered any in the literature.[110] Hence, for the remainder of this project, I operate under the assumption that there is no other, third kind of decision situations.[111]

The second question is whether collective decision situations can be *reduced* to individual ones—whether we *truly* need to recognize a distinct kind of decision situations. This question is far more intriguing and will arise repeatedly throughout this project. Two incompatible claims represent the most compelling possible answers to this question. First, consider:

Claim 5.1 (Compositionalism) All *collective decision situations can be reduced to individual decision situations (plus some structure).*

Note that Compositionalism allows that collective decision situations are more than *just* individual decision situations. Even if reducible to individ-

[110] Another distinction worth considering in this context is the one between decisions made by diachronic persons and those made by their time slices (or temporal parts). A well-known example is the case of Professor Procrastinate (cf. Jackson 2014; Jackson and Pargetter 1986; Woodard 2009), which features prominently in the actualism–possibilism debate in normative ethics (cf. Timmerman and Y. Cohen 2020). This distinction may either 'collapse' into the distinction between individual and collective decision situations—if time slices (or temporal parts) qualify as agents themselves (cf. Dietz 2020)—or it may remain orthogonal to it. In either case, similar structural questions arise, such as whether the decisions of diachronic persons can be reduced to those of their time slices. Given these parallels with the Challenge, I will briefly revisit this debate at the end of this thesis.

[111] As with the different kinds of collective decision situations, there could, of course, also be mixed cases here, i.e., those which are clearly individual decision situations and those which are possibly genuine collective decision situations. But as with other mixed cases, it is enough for us to understand the two pure cases to be able to handle mixed cases well.

ual decision situations, many collective decision situations have to involve *some* structural element, for instance, the temporal order and counterfactual dependencies between earlier and later decisions. Nevertheless, according to Compositionalism, a collective decision situation for n agents involves (at least) $n+1$ decision situations: the collective decision situation and at least one individual decision situation for each agent.[112] Compare Compositionalism to

Claim 5.2 (Genuine Kind View) *Some collective decision situations cannot be reduced to individual decision situations (plus some structure).*

The Genuine Kind View is the negation of Compositionalism. Since they both clearly express meaningful propositions (for example, they are not category mistakes), either one or the other is true. According to the Genuine Kind View, at least some collective decision situations cannot be fully reduced to individual decision situations. Consequently, at least some agents in certain collective decision situations are not merely operating within their 'own' individual decision situations but are also agents within genuine collective decision situations.

The straightforward application of moral theories—which are defined over individual decision situations[113]—is only feasible if Compositionalism holds. Given Compositionalism, we could, in principle, decompose arbitrary collective decision situations into the individual decision situations of the agents involved, enabling us to apply moral theories directly to these individual situations to determine the deontic status of each agent's options. Trivially, this procedure would also align with Methodological Individualism. Recall

Principle 2.1 (Methodological Individualism)
The primary bearers of deontic status are options of moral agents. Whatever has a deontic status and is not itself the option of a moral agent has this status merely in a derivative sense, that is, deontic status is a function of the deontic status of certain options of moral agents.

On the contrary, if the Genuine Kind View were correct, there would be no guarantee that for an arbitrary collective decision situation, there is an individual decision situation for each and every agent within. It is then far from

[112]There might be intermediate or partial collective decision situations involved, for instance, one or several situations of $n-1$ agents, $n-2$ agents and so on.

[113]It could be argued that *I* have defined moral theories in this way, and that this may not reflect a universally agreed understanding of them (if there were such consensus). Nevertheless, I maintain that my definition aligns closely with both everyday language use and the prevailing practices in normative ethics (see Section 2.2, particularly the principle of Methodological Individualism, restated below, as well as the discussion on page 23 and the subsequent pages).

obvious how we should apply moral theories to such irreducible collective decision situations, let alone in a way that is in line with Methodological Individualism (we remind ourselves of Jackson's approach, cf. Subsection 4.4.3). Therefore, determining whether Compositionalism or the Genuine Kind View holds true is of critical importance to this project.

On closer inspection, however, Compositionalism could turn out to be a hopeless position. We have already seen some collective decision situations where at least some arguably morally relevant parts of certain consequences are *not* fully determined by any individual agent's actions. For instance, with respect to our running example Two Factories, even though both Ann and Ben can ensure the pollution of the river (by polluting individually), none of them can ensure that it is *not* polluted. This depends on what *both* agents do. This is to say that apparently some (parts of) consequences are defined *only* over *combinations* of actions. However, if we cannot properly account for these consequences in our moral assessments, we risk—and, in many cases, undoubtedly do—failing to recognize certain morally relevant outcomes of actions. We thus apparently cannot reduce arbitrary collective decision situations to individual ones.[114]

Although I will ultimately seek to defend Compositionalism and, building on it, propose a robust solution to the CHALLENGE, I will first present further evidence *against* Compositionalism. In the next chapter, I will ground my foundational critique of the current understanding of the CHALLENGE on these considerations. Thus, one of the primary objectives of the following sections is to examine key properties of collective decision situations and related distinctions, some of which raise significant doubts about Compositionalism's defensibility. Specifically, we will aim to achieve a precise understanding of the apparent non-decomposability of certain collective decision situations—a characteristic exemplified by Troublemakers, which form a subset of such cases by definition.

5.2 (Semi-)Formalism and Shorthands

As before with individual decision situations, it will prove useful later on to have some shorthands, formal specifications, and notions for collective decision situations and their components. Let there be some collective decision situation D of agents $A_1, \ldots, A_n$ with corresponding option spaces

[114] Some may argue that even such collective decision situations are reducible to individual decision situations, but to situations in which we cannot assign *all* consequences of combinations of actions to individual actions. However, such reductions or decompositions would not be without loss. But since, from the point of view of at least consequentialist theories, essentially important aspects of the collective decision situation would be lost in such 'lossy reductions', this position is unacceptable from the outset, at least for champions of MOAC. In my project, I am concerned solely with the general possibility of *lossless* reductions.

$\Phi_{A_1}, \ldots, \Phi_{A_n}$. Let $\mathcal{A}$ refer to the set of agents in D and let Γ be the set of these agents' option spaces, i.e., $\Gamma := \{ \Phi_A \mid A \in \mathcal{A} \}$. We use Υ to refer to *proper* combinations of actions, possibly with superscripts, i.e., we write $\Upsilon^1, \Upsilon^2, \ldots$ or simply $\Upsilon, \Upsilon', \ldots$ to distinguish between different proper combinations. We write the combination of the actions $\phi_1, \ldots, \phi_n$ as n-tuple $\langle \phi_1, \ldots, \phi_n \rangle$. It is thus natural to use the index notation Υ_i to access the ith element of a combination Υ, i.e., if $\Upsilon = \langle \phi_1, \ldots, \phi_n \rangle$, then $\Upsilon_i = \phi_i$. For convenience, we use the set-theoretic notation of $\phi \in \Upsilon$ to express that there is an index i such that $\phi = \Upsilon_i$.

Next, we turn to the set of proper combinations of D. This set, which is clearly built on top of Γ and that I therefore refer to by Ψ_Γ, I will call the *domain of* D. It turns out that defining Ψ_Γ on a general level is rather complicated because not every combination of actions we can, in principle, construct from an option space is necessarily a proper one, i.e., a combination over which consequences are specified (remember Job Market above). I will come back to some details in a short while.

As always, I occasionally omit unnecessary indexes; for example, I write Ψ for the domain of a collective decision situation instead of Ψ_Γ when the relationship to a specific set of option spaces Γ is either trivial or irrelevant. Similarly, as with individual decision situations, we use $\mathcal{C}$ to refer to the set of relevant contexts of D and define $\mathcal{O} = O_1, \ldots, O_m$ as the set of consequences (or *o*utcomes, hence the shorthand "O") of D, where $m = |\mathcal{C}| \cdot |\Psi|$. Recall that the set of relevant contexts is determined by the relevance stance being considered (cf. Subsection 2.3.1). In the case of MOAC, and in accordance with the Objective View, there is exactly *one* relevant context: the actual one.

To be able to write succinctly about the relationship between combinations and contexts with consequences in $\mathcal{O}$, we again model this relation in terms of an *outcome function* $\mathsf{Out} : \Psi \times \mathcal{C} \to \mathcal{O}$. Sometimes, when more than one decision situation is under consideration, I use indices for disambiguation; where not needed, I leave them out.

As with individual decision situations, we use tuples as abstract representations of collective decision situations: a collective decision D for a set of agents $\mathcal{A}$ with a set of option spaces Γ and an outcome function $\mathsf{Out} : \Psi_\Gamma \times \mathcal{C} \to \mathcal{O}$ is represented by a tuple $\mathsf{D} := \langle \mathcal{A}, \Gamma, \mathsf{Out} : \Psi_\Gamma \times \mathcal{C} \to \mathcal{O} \rangle$ (or, for singletons $\mathcal{C} = \{C\}$, just $\mathsf{D} := \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$). Again, just like with individual decision situations, we refer to a collective decision situation D with a set of relevant contexts $\mathcal{C}$ and a domain Ψ_Γ by just giving one such tuple and, thereby, implicitly introducing both (and also the set of outcomes $\mathcal{O}$) through giving just the signature of the outcome function. Finally, let $\mathfrak{C}$ denote the set of all collective decision situations.

Valuation functions and related concepts can be straightforwardly extended to collective decision situations. Let $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$ be a function that

assigns values from some value space $\mathcal{V}$ to arbitrary outcomes from the now extended

$$\mathcal{W} := \bigcup_{D \in \mathfrak{C}} \mathcal{O}_D \cup \bigcup_{D \in \mathfrak{I}} \mathcal{O}_D.$$

For a set of outcomes $\mathcal{O}_D \subseteq \mathcal{W}$ of some collective decision situation D, let us call $\mathsf{Val}(\mathcal{O}_D) := \{ \mathsf{Val}(O) \mid O \in \mathcal{O}_D \} \subset \mathcal{V}$ the *valuative profile* (of D). Again, we require the existence of a total order $\leq_{\mathcal{O}_D}$ over $\mathsf{Val}(\mathcal{O}_D)$ for arbitrary $D \in (\mathfrak{C} \cup \mathfrak{I})$. Finally, given an individual decision situation D such that $\Upsilon \in \Psi_D$, and a relevant context C, let $\mathsf{Val}_C(\Upsilon)$ be agreed upon as an abbreviated notation for $\mathsf{Val}(\mathsf{Out}_C(\Upsilon))$, exactly as it was defined for the individual case.

Before we continue, let's quickly introduce a more sophisticated version of Collectively Maximizing, much as we did for its individualistic counterpart MOCOR in Part I. Recall

Principle 1.1 (Collectively Maximizing – tentative)
If all agents act rightly, then they are guaranteed to produce the morally best outcome they could together bring about.

Based on the notions established in this section, we can clarify:

Principle 5.1 (Collectively Maximizing) *Let D be a collective decision situation with domain Ψ and with actual context C. If $\Upsilon \in \Psi$ consists only of right actions, then there is (and can be) no alternative $\Upsilon' \in \Psi$ with better consequences than Υ relative to C.*

It is worth digging a bit deeper with respect to proper actions and the domain, even though, for the most part, the combinations within the domain of some collective decision situation, i.e., the proper combinations relative to that situation, should be evident given a sufficiently complete case description. Nevertheless, having a more precise notion of the domain allows for pinning down specific properties that mark different classes of collective decision situations. These classes, in turn, help to make precise restrictions with respect to what kind of collective decision situations I focus on in this project (and *why* I do so). While making these restrictions increases this part's overall readability and comprehensibility, it admittedly comes with a loss of generality. As indicated above, however, this generality can be regained through formal efforts and rigorous, albeit somewhat technical work, which I will spare us here.

5.2.1 Domains and Properties and the Triad

First, let me stress that for now we ignore the possibility of synchronous actions. I will return to that issue later in this part, but for now, we can lay

that possibility, which seems rather esoteric anyway, aside for the time being as it would just complicate matters. For now, we can focus on defining the most general domain possible. This *maximal set of possible combinations* given some set of option spaces Γ I shall refer to as *domain space* Ψ^*_Γ. It serves as a superset from which all actual domains of specific collective decision situations are subsets, i.e., $\Psi_\Gamma \subseteq \Psi^*_\Gamma$. This ensures that any possible combination of actions can be considered relative to a given set of option spaces. Assuming the empty set is *not* a meaningful combination (after all, it must be possible for some choice to be made), the domain space can be formally defined as:

$$\Psi^*_\Gamma = \bigcup_{\tau \in S_n} \bigcup_{i=1}^{n} \bigtimes_{j=1}^{i} \Phi_{A_{\tau(j)}}$$

This definition draws on the concept of *permutations* from combinatorics. Let S_n denote the set of permutations of the (index) set $I_n = \{1, \dots, n\}$, representing all bijections $\tau : I_n \to I_n$. A permutation $\tau \in S_n$ rearranges the elements of I_n, making S_n the set of all possible rearrangements of I_n. In the context of the formula above, these rearrangements enable us to consider every possible ordering of agents and, consequently, all possible sequences of actions. Notably, S_n is also known as the *symmetric group* of I_n,[115] a structure fundamental in combinatorics and group theory.

Collective decision situations with a full domain, i.e., with $\Psi_\Gamma = \Psi^*_\Gamma$ (given their set of option spaces Γ), are not uncommon (think of Two Factories), but also far from being guaranteed, especially in the context of Sequential Cases (recall Job Market). Thus, there still remain a lot of different considerations concerning Ψ_Γ.

I will set Sequential Cases aside for now, limiting my focus to Coordination Cases, before revisiting Sequential Cases at the end of this chapter. For Coordination Cases, I will primarily concentrate on Two Factories–like cases, which exemplify several key properties. Most of these properties translate directly to their domains, which I will assume throughout much of this part of the project.

First, while these cases are *minimal* in that they involve two agents, each with two options, they are *maximal* in the sense that their domain includes only *maximal* combinations, i.e., combinations that contain one action for *each* agent.

Second, the moral quality of the consequences of these combinations is *order-invariant*, i.e., not only is it possible for the agents to act in arbitrary orders, but the outcomes are, in a sense, morally indistinguishable. While this is, strictly speaking, a property concerning the valuative profile of the decision

[115] Where the group operation is the successive application of permutations, and the identity permutation serves as the neutral element.

situations more than of the domain itself, it requires all the corresponding combinations of actions to be proper in the first place.

Third, I focus on *symmetric* Coordination Cases, i.e., in a very rough sense, cases where the quality of the outcome depends only on the performed actions and not on *who* is performing them. All three of these properties will be specified and explicated in more detail below.

I refer to these three properties—maximality, order invariance, and symmetry—collectively as the *Triad*, and I will say that a collective action case exhibiting all three properties *satisfies the Triad*. Except for a few explicitly marked exceptions, the investigations in this part are restricted to Coordination Cases that satisfy the Triad. This subclass of Coordination Cases is both well-populated and encompasses all relevant Mutual Exculpation Cases commonly discussed in the literature. While this restriction entails some loss of generality, it offers significant gains in readability and comprehensibility. Moreover, the cases identified through this approach serve as excellent initial cases for induction-based generalizations, including those extending toward Sequential Cases, as we will later see.

5.2.1.1 Maximality

Maximality is a straightforward property. We can capture maximality formally:

Property 5.1 (Maximality) *Let D be a collective decision situation with a set of agents $\mathcal{A}$ with corresponding option spaces Φ_A for $A \in \mathcal{A}$. A combination of actions is* maximal *(with respect to $\mathcal{A}$) if and only if it contains an action from each agent (from $\mathcal{A}$).*

D is maximal *if and only if every proper combination of actions in its domain is maximal, i.e.*

$$\forall \Upsilon \in \Psi_{\Phi_A} : \exists i : 1 \leq i \leq |\Upsilon| \text{ and } \forall A \in \mathcal{A} : \Upsilon_i \in \Phi_A.$$

For simplicity's sake, let us assume that no agent can perform more than one action (at a time), which is rather plausible in most cases. Against the background of this assumption, we can simplify the formal condition: A collective decision situation is maximal if and only if

$$\forall \Upsilon \in \Psi : |\Upsilon| = |\mathcal{A}|.$$

In terms of the domain, maximality comes down to restricting the domain such that we get (for $\Gamma = \{\Phi_{A_1}, \ldots, \Phi_{A_n}\}$):

$$\Psi^*_{\Gamma,\text{maximal}} \subseteq \bigcup_{\tau \in S_n} \bigtimes_{i=1}^{n} \Phi_{A_{\tau(j)}}.$$

5.2.1.2 Order Invariance

Two Factories exemplifies another property: it is *invariant under the order of actions* (or, more concisely, order-invariant), meaning that only *what* actions the agents perform matters, not the *order* in which they perform them.

As with maximality, we capture order invariance formally. First, we define a shorthand for combinations of actions with the same constitutive actions. For two combinations from the same domain of some decision situation ($\Upsilon, \Upsilon' \in \Psi$), we write:

$$\Upsilon \approx \Upsilon' \quad \text{if and only if} \quad \forall \phi \in \Upsilon : \ \phi \in \Upsilon' \quad \text{and} \quad \forall \phi \in \Upsilon' : \ \phi \in \Upsilon.$$

This definition ensures that $\Upsilon \approx \Upsilon'$ holds when the two combinations contain identical actions, irrespective of order.

Using this shorthand, we define the *set of sequential recombinations* of a given combination Υ, relative to a specific collective decision situation with domain Ψ, as:

$$\overset{\frown}{\Psi}_{\Upsilon} := \left\{ \Upsilon' \in \Psi \mid \Upsilon' \approx \Upsilon \right\}$$

Note that, by construction, $\Upsilon \in \overset{\frown}{\Psi}_{\Upsilon}$.[116]

Based on this notion, we can now define the property of order invariance:

Property 5.2 (Order Invariance) *Let D be a collective decision situation with domain Ψ. D is* invariant under the order of action *(or, shorter,* order-invariant*) if and only if, for every proper combination of actions, all its sequential recombinations are also proper, and the outcomes of all these combinations are valuatively equivalent. Formally:*

$$\forall\, \Upsilon \in \Psi : \ \overset{\frown}{\Psi}_{\Upsilon} \subseteq \Psi \textit{ and } \forall\, \Upsilon' \in \overset{\frown}{\Psi}_{\Upsilon} : \ \mathsf{Val}(\mathsf{Out}(\Upsilon)) = \mathsf{Val}(\mathsf{Out}(\Upsilon')).$$

Both being maximal and order-invariant are typical properties of Coordination Cases. While I indeed restrict this investigation to Coordination Cases with these properties, not all Coordination Cases *necessarily* have these properties.[117]

[116] This inclusion holds because $\Upsilon \approx \Upsilon$ is trivially true for any combination.

[117] An example without these properties can be easily construed, at least if we allow a certain kind of case distinction. Consider

Case 5.1 *Mutual Termination John has kidnapped Adam and Lawrence. Lawrence wakes up in a small room with a timer showing 10 minutes and a big red button. Adam wakes up simultaneously in another room with the same setup. If both wait ten minutes without pressing their button, the doors open, and both stay alive. However, they also receive no additional reward. If Adam pushes his button first, he gets free and gets one million dollars. However, this floods Lawrence's room with a poisonous gas that kills him cruelly within seconds. In contrast, if Lawrence pushes first, he kills Adam through toxic gas and gets free, but Adam would not get*

At this point, it is useful to introduce the notions of an *equivalence relation* and *equivalence classes* for combinations of actions, relative to a decision situation and a valuation function. These concepts will prove valuable later in our analysis. Let D be a collective decision situation with $\mathsf{D} := \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$, and let Val be a valuation function. Then, for combinations $\Upsilon, \Upsilon' \in \Psi$, the relation $\sim \subseteq \Psi \times \Psi$ is defined as:

$$\Upsilon \sim \Upsilon' \quad \text{if and only if} \quad \mathsf{Val}(\mathsf{Out}_C(\Upsilon)) = \mathsf{Val}(\mathsf{Out}_C(\Upsilon')),$$

This is clearly an equivalence relation.[118] Using this relation, we define the equivalence class of a combination as:

$$[\Upsilon] := \{ \Upsilon' \in \Psi \mid \Upsilon \sim \Upsilon' \}.$$

These concepts connect naturally to the property of order invariance. Specifically, in any order-invariant decision situation, the following holds:

$$\forall \Upsilon, \Upsilon' \in \Psi : \Upsilon \approx \Upsilon' \to \Upsilon \sim \Upsilon'.$$

Alternatively, this relationship can be expressed in terms of the introduced classes:

$$\forall \Upsilon, \Upsilon' \in \Psi : \Upsilon' \in \overset{\curvearrowleft}{\Psi}_\Upsilon \to \Upsilon' \in [\Upsilon].$$

In simpler terms, if two proper combinations of actions contain the same constitutive actions, they will yield outcomes of equal value.

---

any reward. If Adam pushes first and Lawrence, in his death throes, also *pushes his button, not only will Adam also die, but John will next kidnap Adam's wife for one of his cruel experiments; but if Lawrence pushes first and Adam, in his death throes, also presses his button, this has* no *additional effect (beyond Adam's death). Similarly, in the improbable case of actual synchronous pushing, both die.*

Here is a normal form representing the case:

		Lawrence	
		not-push	push
Adam	not-push	both live on, no rewards (best)	only Lawrence lives on, but no reward (second-worst?)
	push	only Adam lives on and is rich (second-best)	both dead, but if Adam pushed first, his wife is kidnapped next (worst)

It is easy to see that Mutual Termination is not maximal (because 'combinations' of one agent's action are proper since not-pushing is by description meant to be a deliberate action, but one that is not necessary to 'resolve' the situation). Nor is it order-invariant (in the case where both push their buttons, it becomes important—for Adam's wife, at least—whether Adam pushed first).

[118] This follows trivially because = is an equivalence relation over the value space.

Based on this notion, we can introduce a more fine-grained conception of equivalence that considers both the values of the outcomes and the combination-constituting actions. This refined equivalence relation will later prove useful for reducing the state space of *generalized extensive forms*, a unifying representation for Coordination Cases and Sequential Cases. Let D be a collective decision situation defined as $\mathsf{D} := \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$, and let Val be a valuation function. Then the relation $\sim^* \subseteq \Psi \times \Psi$ is defined, for two combinations $\Upsilon, \Upsilon' \in \Psi$, as follows:

$$\Upsilon \sim^* \Upsilon'$$

if and only if

$$\mathsf{Val}(\mathsf{Out}_C(\Upsilon)) = \mathsf{Val}(\mathsf{Out}_C(\Upsilon')) \wedge \big(\forall \phi \in \Upsilon : \phi \in \Upsilon'\big) \wedge \big(\forall \phi \in \Upsilon' : \phi \in \Upsilon\big)$$

This relation is, again, an equivalence relation.[119] Further, we define the equivalence class of a combination under this refined relation as:

$$[\Upsilon]^* := \{ \Upsilon' \in \Psi \mid \Upsilon \sim^* \Upsilon' \}.$$

By construction, $[\cdot]^*$ further refines the earlier $[\cdot]$ equivalence classes by requiring that combinations belong to the same class *only if* they yield morally equivalent outcomes *and* involve the same individual actions.

Next, we define the set of equivalence classes for a given collective decision situation D with domain Ψ (relative to some valuation function) as:

$$[\Psi]^* := \{ [\Upsilon]^* \mid \Upsilon \in \Psi \}$$

and correspondingly, we define the *set of sets of representatives* as:

$$\mathcal{R} := \{ R \subseteq \Psi \mid \forall [\Upsilon]^* \in [\Psi]^* : \exists! \Upsilon \in R : \Upsilon \in [\Upsilon]^* \}.$$

In other words, every $R \in \mathcal{R}$ is a set that, for each equivalence class $[\cdot]^*$ of combinations relative to D (and an arbitrary but fixed valuation function), contains *exactly one* element from that class. We can imagine constructing such a *set of representatives* $R \in \mathcal{R}$ by iterating through all equivalence classes in $[\Psi]^*$ and picking precisely one element from each class.

Now, we come to the final crucial property for this investigation.

5.2.1.3 Symmetry

The third property central to this part of my project is *symmetry*. Symmetry is a property of Coordination Cases.[120] Intuitively, a Coordination Case is

[119] The reflexivity, symmetry, and transitivity of $\sim^*$ follow directly from the fact that $=$ is an equivalence relation over the value space and that the universal quantifier together with the $\in$ relation preserves these properties. A formal proof is left as an exercise for the reader.

[120] While symmetry can also be generalized to Sequential Cases, exploring this generalization lies outside the scope of this thesis. For now, I focus solely on its role in Coordination Cases.

symmetric if and only if the specific identities of the agents performing actions are irrelevant to the outcome; as long as agents perform *corresponding* actions, the outcomes remain identical with respect to their morally relevant qualities.

While the notion of *corresponding actions* can make formalization challenging, it is often clear in practice. For instance, Two Factories is a symmetric case: whether Ann pollutes and Ben produces cleanly or Ben pollutes and Ann produces cleanly, the outcomes are *qualitatively* identical by assumption. Although the specific workers affected may differ, the harm done to them is assumed to be identical in nature and magnitude. Here, the correspondence between actions is implied by their type (e.g., polluting vs. clean production).[121]

However, if we would—for whatsoever reason—think of Ann's option of polluting as corresponding to Ben's option of producing cleanly, and vice versa, we might think that the situation was *not* symmetrical. To illustrate this, let us assume that we number the agents' options and give corresponding options the same number. If we then assume that in the normal form, we list the options of the agents according to their number in ascending order—from top to bottom and from left to right—, the intuitive correspondence mapping of Two Factories would give our well-known normal form

		Ben	
		pollute	produce cleanly
Ann	pollute	second-worst	worst
	produce cleanly	worst	best

while the unintuitive mapping would give us this one:

		Ben	
		produce cleanly	pollute
Ann	pollute	worst	second-worst
	produce cleanly	best	worst

Obviously, the first normal form is symmetrical in a *geometric* sense: the normal form could be 'mirrored' along the main diagonal, i.e., along the imaginary line from the upper left to the top right, without changing the normal form in a morally relevant way (at least from a consequentialist point of view).

[121] Many well-known examples share this structure, likely because it simplifies description. However, it is by no means a necessary condition for Troublemaker, as I will demonstrate with a more complex case later in this section. This likely explains why Kagan, in several instances, referred to actions "of a given kind" (cf. Subsection 4.4.1).

This is not true for the second normal form because the consequences of both polluting are worse than those of both producing cleanly.

While in Two Factories the 'correct' representation may appear obvious—guided, so to speak, by the agents' option types—the following example demonstrates that this is not always the case. Consider

Case 5.2 (Lucky Lisa) *Johnny is a doctor, and he is really bad at his job. When Lisa asked him for some painkillers against her migraine, he accidentally gave her drug X, which, under normal circumstances, would cause Lisa tremendous pain, pain much worse than her migraine. However, drug X also makes her immune against truth serum Y for several hours. At the same time, getting injected with that serum would also neutralize the painful effects of drug X.*

After leaving Johnny's practice, Lisa takes her dose of drug X on her way home, where Mark is already waiting for her, hiding behind her front door. Mark wants to squeeze some secret information out of Lisa with the help of truth serum Y. This information would allow Mark to blackmail Lisa in the future, inflicting as much disutility on her as drug X's default pain effect does. After overpowering Lisa and tying her up to a kitchen chair, he injects truth serum Y into her. Thanks to drug X, Lisa can keep her secret and lie to Mark so that he ultimately fails with his evil plans to blackmail Lisa and, at the same time and unintendedly, he neutralizes drug X's painful effect. Still, Lisa suffered from a migraine for the rest of the day.

One way to represent Lucky Lisa in a normal form is this

		Johnny	
		give drug X	give pain killers
Mark	inject serum Y	second-worst	worst
	don't inject serum Y	worst	best

Obviously, our two protagonists, Johnny and Mark, are part of a Trouble-maker that is structurally equivalent to Two Factories. This normal form is symmetric for the same reasons. However, consider this alternative representation of Lucky Lisa:

		Johnny	
		give pain killers	give drug X
Mark	inject serum Y	worst	second-worst
	don't inject serum Y	best	worst

Geometrically speaking, this form is *not* symmetrical. However, unlike with Two Factories, I cannot really see why one of the two normal forms should be more representative than the other. I think they are on a par.

The right reaction to this observation is not to introduce a convoluted theory of correct or adequate correspondence relations. I won't take a stand on the question of what makes correspondence mappings adequate or whether this idea even makes sense in a general setting or proves theoretically fruitful. Instead, let us call all decision situations symmetric for which there is at least one symmetric normal form. Here is a more precise definition:[122]

Property 5.3 (Symmetry) *Let D be a maximal collective decision situation with domain Ψ and two agents A_1 and A_2 with corresponding option spaces*

$$\Phi_{A_i} = \left\{ \phi^1_{A_i}, \phi^2_{A_i} \right\}$$

for $i \in \{1,2\}$. D is symmetric *if and only if there is at least one mapping between the agents' option spaces such that, for every proper combination of actions, the outcomes are valuatively equivalent to the combination that results from the original combination by applying that mapping, i.e., there is a permutation $\tau : I_2 \to I_2$ (i.e., $\tau \in S_2$) such that, for $i,j,k,l \in \{1,2\}$ and $k \neq l$ it holds that*

$$\forall \left\langle \phi^i_{A_k}, \phi^j_{A_l} \right\rangle \in \Psi : \mathsf{Val}\Big(\left\langle \phi^i_{A_k}, \phi^j_{A_l} \right\rangle \Big) = \mathsf{Val}\Big(\left\langle \phi^{\tau(j)}_{A_k}, \phi^{\tau(i)}_{A_l} \right\rangle \Big).$$

In other words, symmetry ensures that there exists a correspondence between the actions of the agents such that the outcomes remain valuatively equivalent regardless of which agent performs which action. This means that for any proper combination of actions, we can reorder or permute the actions between agents without altering the moral quality of the resulting outcome, i.e., such that they line up to a symmetric normal form.

It is easy to mix up order invariance and symmetry. To highlight the difference between the two, consider the following case which is order-invariant but not symmetric:

Case 5.3 (Henry's Hardship) *In the aftermath of a severe recession, Henry has lost his apartment and is now homeless. His friend Molly has a vacant, albeit unfurnished, room in her apartment that she could offer to Henry free of charge. Meanwhile, Rico, who has also been struggling, lives in a cramped apartment overflowing with excess furniture he no longer needs.*

If Rico gives Henry the furniture but Molly does not offer him the room, the gesture provides little help and even worsen Henry's situation: either Henry accepts the furniture, leaving him homeless while also burdened with cumbersome possessions, or he rejects Rico's offer, angering him.

In contrast, if Molly offers Henry the room, he is better off regardless of Rico's actions. Naturally, Henry would be in the best possible situation if he were to receive both the room and *the furniture.*

[122] For simplicity's sake, I restrict it to maximal Coordination Cases involving two agents with the same number of options.

		Rico	
		give furniture away	don't give furniture away
Molly	offer room	best	second-best
	don't offer	worst	second-worst

		Rico	
		don't give furniture away	give furniture away
Molly	offer room	second-best	best
	don't offer	second-worst	worst

		Rico	
		give furniture away	don't give furniture away
Molly	don't offer	worst	second-worst
	offer room	best	second-best

		Rico	
		don't give furniture away	give furniture away
Molly	don't offer	second-worst	worst
	offer room	second-best	best

Table 5.1: The four normal forms for Henry's Hardship. Note that there are four normal forms instead of the expected two (as suggested by Property 5.3, given that $|S_2| = 2$). This is because there are two possible ways to enumerate the first agent's options, resulting in two sets of two normal forms each. When considered crosswise (at the meta-level), the corresponding normal forms are strictly equivalent in terms of symmetry, as the same options appear on the main diagonal. Thus, Property 5.3 is sufficiently general to encompass all possible cases.

This case can be represented in terms of four different normal forms (cf. Table 5.1). Obviously, none of these normal forms is symmetric, and hence Henry's Hardship is not symmetrical. Yet, the case, like all Coordination Cases, is order-invariant.

Two Factories, however, is both order-invariant and symmetric, and these two properties together create a structure that deserves closer examination. That Two Factories is *order-invariant* means, among other similar equalities, that

$$\langle \text{pollute}_{\text{Ann}},\ \text{produce cleanly}_{\text{Ben}} \rangle$$

has valuatively identical outcomes as

$$\langle \text{produce cleanly}_{\text{Ben}},\ \text{pollute}_{\text{Ann}} \rangle .$$

However, that Two Factories is *symmetric* means that

$$\langle \text{pollute}_{\text{Ann}},\ \text{produce cleanly}_{\text{Ben}} \rangle$$

has valuatively identical outcomes as

$$\langle \text{produce cleanly}_{\text{Ann}},\ \text{pollute}_{\text{Ben}} \rangle .$$

As we have seen in the first part, the typically discussed Troublemakers that are Coordination Cases are maximal, order-invariant, and symmetric, i.e., cases satisfying the Triad. Accordingly, as announced, I restrict my investigation primarily to Coordination Cases with these properties. Interestingly,

for every 'triadic' Coordination Case with domain Ψ which is symmetric according to some correspondence mapping τ it holds that for any combination $\left\langle \phi^i_{A_1}, \phi^j_{A_2} \right\rangle \in \Psi$:

$$\begin{array}{ccc} \mathsf{Val}\left(\left\langle \phi^i_{A_1}, \phi^j_{A_2} \right\rangle\right) & = & \mathsf{Val}\left(\left\langle \phi^{\tau(i)}_{A_1}), \phi^{\tau(j)}_{A_2} \right\rangle\right) \\ \shortparallel & & \shortparallel \\ \mathsf{Val}\left(\left\langle \phi^j_{A_2}, \phi^i_{A_1} \right\rangle\right) & = & \mathsf{Val}\left(\left\langle \phi^{\tau(j)}_{A_2}), \phi^{\tau(i)}_{A_1} \right\rangle\right) \end{array}$$

The horizontal equalities are guaranteed by symmetry, while the vertical equalities are warranted by order invariance. That all these combinations are proper is given by maximality.

5.2.2 1-Variants and Independence

Finally, note that every maximal Coordination Case is, by definition, act-independent in the sense of Independency of Action as introduced earlier. Recall:

Property 3.9 (Independency of Action) *Let $\mathcal{D}$ be a collective decision situation. A combination of actions that is proper within $\mathcal{D}$ is* act-independent *if and only if any combination resulting from a unilateral deviation from that original combination is also proper within $\mathcal{D}$. The decision situation $\mathcal{D}$ itself is called* act-independent *if and only if all combinations of actions in $\mathcal{D}$ are act-independent.*

After all, for every combination Υ in such domain Ψ, we can substitute any action of any agent with another option of that agent without leaving the domain, i.e., we get another combination that is also proper. The following notion captures this operation of unilateral modification of combinations:

Definition 5.2 (1-Variant of a Combination)
Let $\mathsf{D} := \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O} \rangle$ be a collective decision situation. Let $\Upsilon \in \Psi$ be an arbitrary proper combination of actions. Υ' is a 1-variation *of Υ if and only if Υ' differs from Υ with respect to exactly one action of one agent.*

We can now restate the above thought: Given that a case is maximal, we can be sure that every 1-variant of a proper combination within that case is also proper. Without the assumption of maximality and, thus, Independency of Action, the relevant notion—the idea of *minimal* variants of combinations—becomes much harder to capture but allows for much more sophisticated acceptable dependencies of actions in Troublemakers, i.e., for a higher level of generality.[123]

[123] This was already indicated in footnote 91. I leave these details for another occasion.

Based on the notion of 1-variants, we can straightforwardly define the *set of 1-variants* $\Psi_\Gamma^{\Upsilon,1}$ for an arbitrary combination of n actions $\Upsilon \in \Psi_\Gamma$ relative to a domain Ψ_Γ over a set of option spaces Γ. We define

$$\Psi_\Gamma^{\Upsilon,1} := \left\{ \Upsilon' \in \Psi_\Gamma \mid \exists! \, \phi_A \in \Upsilon : \, \phi_A \notin \Upsilon' \wedge \exists! \, \phi'_A \in \Phi_A : \, \phi'_A \in \Upsilon' \right\}.$$

This notion will make capturing the formal structure of Troublemakers later quite easy. Again, when obvious, we will drop the sometimes unnecessary index Γ and will simply write $\Psi^{\Upsilon,1}$.

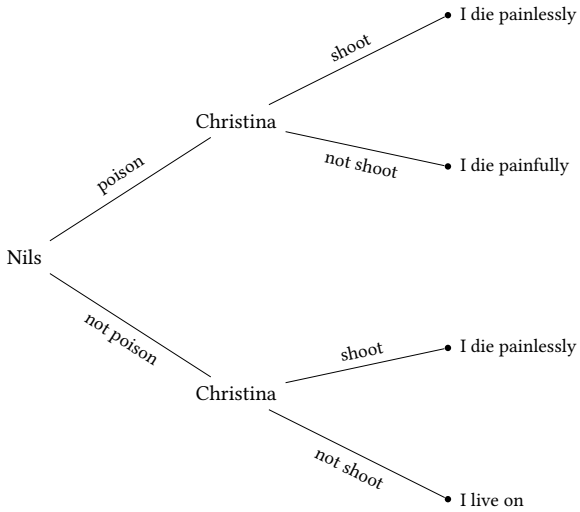
5.3 Revisiting Sequential Cases

While until now I restricted to Coordination Cases, it is finally time to turn to Sequential Cases again. Since Sequential Cases are defined as collective decision situations in which the order is specified, it cannot surprise that Sequential Cases are normally *not* order-invariant. Consider Job Market. The combination of Paul accepting the job first and George declining it second makes no sense, given the case description. The case is also not maximal: George's action of accepting the job alone suffices to specify an outcome.

Obviously, a Sequential Case *can* be maximal. Consider the following case[124] inspired by Derek Parfit (1984, p.70):

Case 5.4 (Deadly Evening) *Nils could trick me into drinking poison of a kind that causes a painful death within a few minutes. Before this poison takes effect, Christina can kill me painlessly by shooting me.*

The extensive form of Deadly Evening is given by



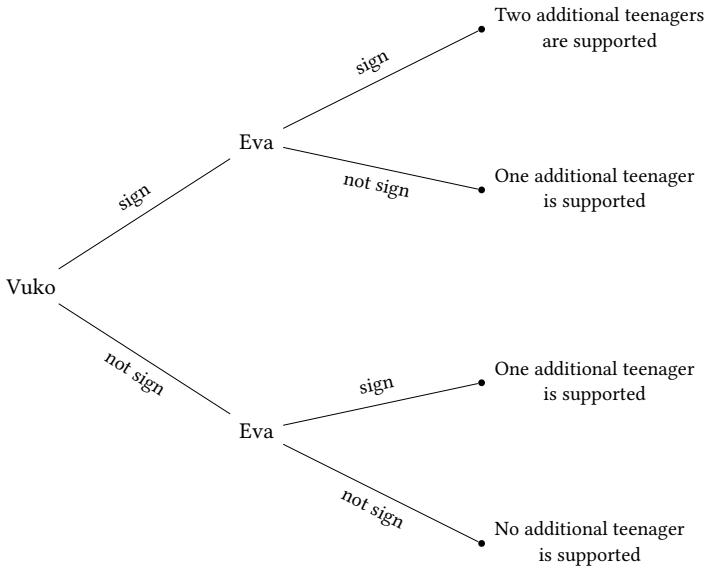
[124] We can also refer to the modified version of Job Market on page 56, where I have already indicated the following case.

Deadly Evening, no doubt, is an exemplar of a *maximal* Sequential Case (and even a Threshold Case with threshold size of one).

Similarly, Sequential Cases can *even* be order-invariant, at least in a slightly different sense than introduced above. Consider, for example, this case:

Case 5.5 (Charity) *Tina collects signatures for a charity project and goes from house to house. For each signature, the city sponsors one socially disadvantaged teenager with a scholarship. Vuko and Eva live in the last two houses on Tina's tour.*

Charity can be represented by the following extensive form:



One could argue that Charity is order-invariant insofar the outcomes would be equally good even if Tina first rang at Eva's door and then at Vuko's, no matter what both decide to do. However, as the order of decisions is fixed to first Vuko, then Eva, these alternative order combinations are improper. Further, Sequential Cases are, in this counterfactual sense, order-invariant are 'boring', as they can be (losslessly) separated into two qualitatively identical, independent individual decision situations—which will be demonstrated in a moment. So, while order invariance is not a truly interesting property with respect to Sequential Cases, at least in the context of the CHALLENGE, the question of what makes a collective decision situation *decomposable* in that sense is crucial for this project.

5.4 Decomposability, Reduction, Conditionalization

There are at least two approaches to "deriving" individual decision situations from collective ones: the first, which I will term *decomposition*, and the second,

reduction (through conditionalization). While decomposition may not be applicable to all collective decision situations, *some* appear to be decomposable. This observation lends support to the Genuine Kind View and challenges Compositionalism. To clarify, let us recall these two conflicting views:

Claim 5.1 (Compositionalism) All *collective decision situations can be reduced to individual decision situations (plus some structure).*

Claim 5.2 (Genuine Kind View) *Some collective decision situations cannot be reduced to individual decision situations (plus some structure).*

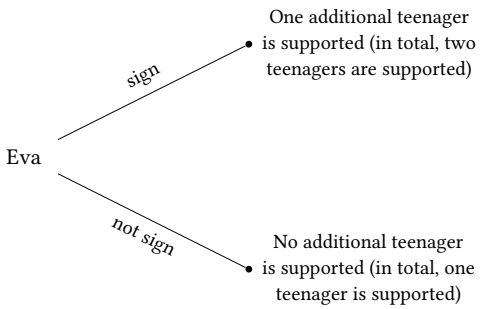
I shall refer to collective decision situations that permit *decomposition* as *decomposable*. I claimed above that Charity is an example of a decomposable collective decision situation. Indeed, it can be decomposed into individual decision situations without losing morally relevant information, at least from a consequentialist perspective: we can assess Eva's decision situation independently from Vuko's action. Deadly Evening, however, is an example of a *non-decomposable* decision situation: we cannot assess Christina's decision independently from Nils' action. These claims certainly need some support, and I will use the technique of *conditionalization* as introduced in Chapter 3. Recall:

Principle 3.10 (Conditionalization) *Let C be a context, and let F be some state of affairs. If it is true, relative to $[\![C \oplus F]\!]$, that p, then it is true, relative to context C, that [if f, then p].*

Recall that "$[\![C \oplus F]\!]$" is shorthand for "C, combined with the assumption that F obtains". In other words, it indicates that, given C as the actual context of a decision situation D, the circumstances of D are *expanded* by incorporating F. Let us illustrate this concept by applying it to Charity. Assume:

Fact 5.1 *Vuko signs the petition.*

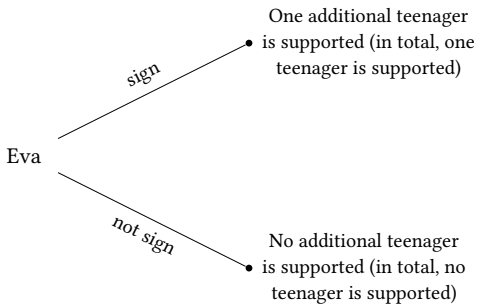
If we combine the context implied by the description of Charity with Fact 5.1, then arguably, there remains an individual decision situation for Eva, namely:



This is what I call a *structural reduction of a collective decision situation by conditionalizing on some action*—in this specific case, a reduction of Charity by conditionalizing on the action referred to in Fact 5.1. Alternatively, we can conditionalize on

Fact 5.2 *Vuko does* not *sign the petition.*

Then we get

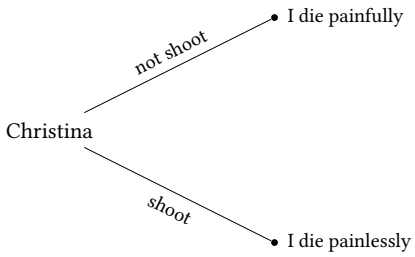


Let's assume that each supported teenager benefits equally from the scholarship. Then, the gain in value that Eva can contribute by signing (or that Eva can block by refusing to sign, respectively) is the same in both cases. This means that what Eva can contribute, morally speaking, is *independent* from what Vuko does. Accordingly, we can assess Eva's situation without considering Vuko's action. That is, we can reduce Charity into two valuatively identical individual decision situations for Eva, resulting from conditionalizing on Vuko's two actions. Note that the same is true *vice versa*: we can assess Vuko's decision without considering Eva's action. For that reason, there is no particular reason to consider Charity in its entirety, as we could also consider the decision situations of both agents separately without losing anything.

Compare this with Deadly Evening. First, consider

Fact 5.3 *Nils poisons me.*

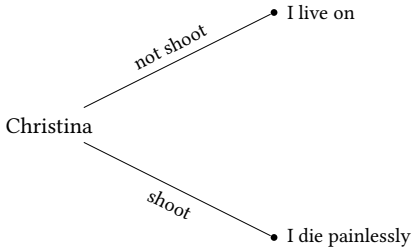
This reduces Deadly Evening to



In this case, arguably, it is better for me if Christina shoots me. But now consider:

Fact 5.4 *Nils does* not *poison me.*

This leaves Christina with



There is no doubt that I am far better off in this case if Christina does *not* shoot me. Therefore, in moral terms, Christina's decision situation depends crucially on what Nils does. Accordingly, we cannot assess her situation without considering Nils' actions. This difference is what makes situations like Charity *decomposable* but Deadly Evening not. We can extract individual decision situations in the described sense, which ultimately allows us to reduce the collective decision situation that we can assess in isolation by applying MOCOR. For non-decomposable situations like Deadly Evening, this cannot be done. All we get are conditional assessments by applying Conditionalization.

This distinction is related to the so-called Sure-Thing Principle.[125] Translated to the moral, consequentialist domain, the Sure-Thing Principle suggests, roughly, that when evaluating potential outcomes relative to certain indeterminacies, these indeterminacies can be disregarded if the outcomes are occurring regardless of how those indeterminacies unfold. We can capture the Sure-Thing Principle in similar terms as we did with Conditionalization:

Principle 5.2 (Sure-Thing Principle) *Let C be a context, and let F be some fact that will either obtain or not (which we refer to as $\neg F$). If it is true, relative to $[\![C \oplus F]\!]$, that p, and it is true, relative to $[\![C \oplus \neg F]\!]$, that p, then it is true, relative to C, that p.*

We can now state the distinction between decomposable and non-decomposable collective decision situations like this:

[125] The *locus classicus* for the Sure-Thing Principle is certainly L. J. Savage (1954). Richard Jeffrey (1982) and, much more recently, Judea Pearl (2000) highlighted the importance of certain independence assumptions between cause/action and effect given the background model. However, since we have agreed to presuppose this type of independence in the cases under consideration, it need not concern us. The Sure-Thing Principle has been contested by Colin Blyth (1972), but his attack based on Simpson's Paradox (see Simpson 1951) violates the independence assumption as well (see Pearl 2016).

Definition 5.3 (Decomposability) *Let D be a collective decision situation of n agents. D is* decomposable *if and only if Conditionalization allows us to reduce D into a class of evaluatively identical individual decision situations for each agent.*

Note that the Sure-Thing Principle gives us good reasons to adopt Decomposability. Because, if a collective decision situation is decomposable, the Sure-Thing Principle allows us to infer non-conditional assessments for all n agents within that situation. However, the other direction is not true: Not every situation allowing unconditional assessments is decomposable in that sense. After all, what needs to be constant over the results of conditionalizations in order to allow for unconditional assessments is only the *ranking* of outcomes, but not the exact valuative profile (as demanded by Decomposability). Recall Henry's Hardship and the corresponding normal form:

		Rico	
		give furniture away	don't give it away
Molly	offer room	best	second-best
	don't offer	worst	second-worst

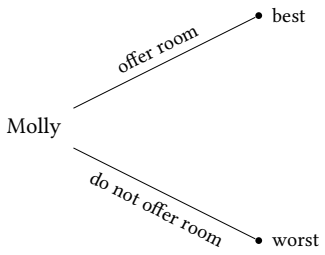
Trivially, Conditionalization and the Sure-Thing Principle can be applied also to Coordination Cases. For instance, relative to Henry's Hardship it is true that

(25) It is better if Molly offers the free room to Henry

because no matter what Rico does or will do, it is better for Henry not to be homeless. But the outcomes still differ morally. Conditionalizing on

Fact 5.5 *Rico gives his spare furniture to Henry.*

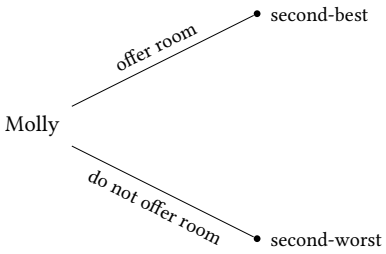
leaves Molly with the individual decision situation



while conditionalizing on

Fact 5.6 *Rico does* not *give his spare furniture to Henry.*

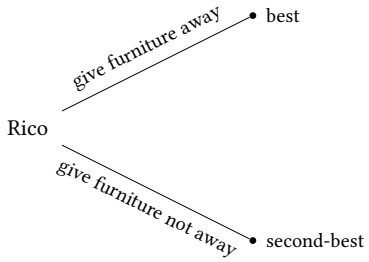
leaves Molly with



However, for Rico, we only get conditional assessments by conditionalizing on Molly's potential actions. Because conditionalizing on

Fact 5.7 *Molly offers her room to Henry.*

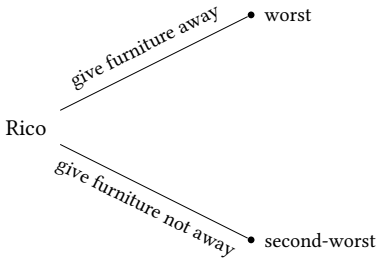
leaves Rico with the individual decision situation



while conditionalizing on

Fact 5.8 *Molly does not offer her room to Henry.*

leaves Rico with



In light of Fact 5.7, it would be right for Rico to give Henry his spare furniture; however, it is wrong for him to do so in light of Fact 5.8. For Rico, we can only get conditional assessments by applying Conditionalization.

The example illustrates three key points. First, conditionalizing is applicable not only to Sequential Cases but also to Coordination Cases. Second, the possibility of arriving at unconditional assessments in a collective decision situation with the Sure-Thing Principle does *not* imply that the situation is decomposable. Sometimes, it is enough that the moral ranking of the outcomes is good enough, as is the case for Henry's Hardship with respect to Molly. Such cases are to some extent uninteresting for this project, although they are *not* decomposable in the strict sense. Reducing them implies losing some morally relevant information, even if not enough to affect moral assessments in terms of MOAC. Third, a collective decision situation may be uninteresting with respect to one agent and interesting with respect to another precisely because it does not allow unconditional assessments for that agent. This is the case, for example, with respect to Rico in Henry's Hardship.

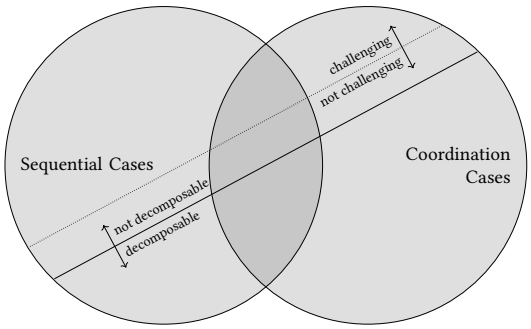


Figure 5.3: The relation between the different sorts of collective decision situations. There are challenging (non-decomposable) and non-challenging (decomposable) cases of both kinds. Furthermore, all challenging cases are non-decomposable, but there are non-challenging cases that are non-decomposable, e.g., Henry's Hardship.

Summing up, decomposable situations are thus only a specific class of what rightfully might be called *non-challenging* (collective decision) situations, i.e., of collective decision situations that pose no particular challenge to MOAC. With respect to the Compositionalism versus Genuine Kind View question we started this section with, decomposable cases do not matter. Figure 5.3 gives an overview of the different kinds of cases introduced so far.

5.4.1 A Formal Toolbox for Reductions

For the remainder of this part, having clear semantics for conditional assessments proves very useful. Thus, we ensure that reductions can be formally captured in the proposed collective decision-making framework. In order to do so, we first define the following operator:

$$\Upsilon \ominus \phi := \begin{cases} \langle \psi_1, \ldots, \psi_{i-1}, \psi_{i+1}, \ldots, \psi_n \rangle, & \text{if } \Upsilon = \langle \psi_1, \ldots, \psi_{i-1}, \phi, \psi_{i+1}, \ldots, \psi_n \rangle, \\ \Upsilon, & \text{otherwise.} \end{cases}$$

In other words, the operator $\ominus$ removes an option ϕ from a combination if the option is part of the combination. If not, it leaves the combination unaffected.

Equipped with $\ominus$, it is easy to neatly define the reduction of collective decision situations, which we have already casually practiced many times in the last section. Let D be some collective decision situation with $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$. We write $D_{\downarrow\phi}$ for the reduction of D by conditionalizing on some option ϕ from some option space Φ from Γ corresponding to the options of an agent $A \in \mathcal{A}$. Then $\mathsf{D}_{\downarrow\phi}$ is defined as

$$\mathsf{D}_{\downarrow\phi} := \left\langle \mathcal{A}_{\downarrow\phi}, \Gamma_{\downarrow\phi}, \mathsf{Out}_{[\![C \oplus \phi]\!]} : \Psi_{\Gamma_{\downarrow\phi}} \to \mathcal{O} \right\rangle$$

with

$$\begin{aligned} \mathcal{A}_{\downarrow\phi} &:= \mathcal{A} \setminus \{A\}, \\ \Gamma_{\downarrow\phi} &:= \Gamma \setminus \{\Phi\}, \\ \Psi_{\Gamma_{\downarrow\phi}} &:= \{\, \Upsilon \ominus \phi \mid \Upsilon \in \Psi_\Gamma \text{ with } \phi \in \Upsilon \,\}, \\ \mathsf{Out}_{[\![C \oplus \phi]\!]}(\Upsilon_{\downarrow\phi}) &:= \mathsf{Out}_C(\Upsilon) \text{ with } \Upsilon_{\downarrow\phi} := \Upsilon \ominus \phi. \end{aligned}$$

We can iterate this operator for collective decision situations with more than two agents. This allows us to reduce *arbitrary* collective decision situations of n agents by conditionalizing on the actions of $n-1$ agents to an individual decision situation of the remaining agent. For this, we first define a *proper part* of a combination of actions Υ to be a combination of actions Υ' —in symbols: $\Upsilon' \sqsubset \Upsilon$—such that $|\Upsilon'| < |\Upsilon|$ and $\forall \phi \in \Upsilon' : \phi \in \Upsilon$. For two combinations with $\Upsilon' \sqsubset \Upsilon$ we define $\Upsilon \ominus \Upsilon'$ to be the *missing part* of (or the *difference* between) Υ with respect to Υ', i.e., if $\Upsilon \ominus \Upsilon' = \Upsilon''$, then $\Upsilon' \oplus \Upsilon'' = \Upsilon$.

Now we can define $D_{\downarrow\Upsilon'}$ for the reduction of D by conditionalizing on some proper part Υ' of a proper combination $\Upsilon \in \Psi$ of D as the repeated application of the above-defined reduction in the correct order:

$$D_{\downarrow\Upsilon'} := \left(\ldots \left(D_{\downarrow\Upsilon'_1} \right)_{\downarrow\Upsilon'_2} \ldots \right)_{\downarrow\Upsilon'_{|\Upsilon'|}}$$

This implies that we can reduce every collective decision situation (with a finite number of agents) to some *individual* decision situation, namely by reducing for a proper part Υ' of a proper combination Υ with $|\Upsilon'| = |\Upsilon| - 1$.

This allows me to extend my definition of the rightness property to finally 'unlock' conditional assessments formally. That is, we want to define what it means, given a *collective* decision situation D, to say that

(26) If the partial combination of actions Υ' were realized, then it is right to ϕ for A

or, somewhat more naturally, given that $\Upsilon' = \langle \phi_{A_{i_1}}, \ldots, \phi_{A_{i_{n-1}}} \rangle$:

(27) If A_{i_1} performs $\phi_{A_{i_1}}$ and ... and $A_{i_{n-1}}$ performs $\phi_{A_{i_{n-1}}}$, then it is right to ϕ for A.

So, let $T, D, C \vDash R(\ \phi \mid \Upsilon'\ )$ express proposition (26) and let D be a collective decision situation with actual context C. We define:

$$T, D, C \vDash R(\ \phi \mid \Upsilon'\ ) \quad \text{if and only if} \quad T, D_{\downarrow \Upsilon'}, [\![C \oplus \Upsilon']\!] \vDash R\phi.$$

This allows us to express the Sure-Thing Principle, restricted for simplicity to a case with two agents, A_1 and A_2, each with two options, ϕ_i and $\neg\phi_i$ (for $i \in \{1, 2\}$): if $T, D, C \vDash R(\ \phi_1 \mid \phi_2\ )$ and $T, D, C \vDash R(\ \phi_1 \mid \neg\phi_2\ )$, then $T, D, C \vDash R\ \phi_1$. The same applies analogously for $\neg\phi_1$, ϕ_2, and $\neg\phi_2$.

Applying all this to our running example, Two Factories, we thus can decide that the intuitive assessments of MOCOR, which played a central role in the reconstruction in Subsubsection 3.5.2.3, really have a solid basis in carefully applied theory. For instance,

(28) If Ben pollutes, it is right for Ann to pollute

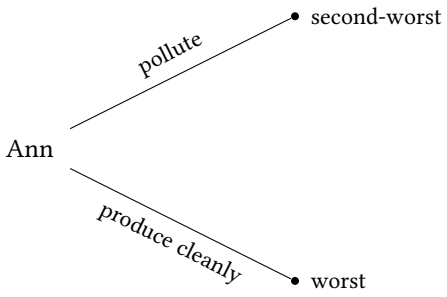
is true in Two Factories because

(29) It is right for Ann to pollute

is true in Two Factories if combined with the fact that

Fact 5.9 *Ben pollutes*

which results in the individual decision situation



After all, in this situation, the best that Ann can do is to pollute.

We now have a solid foundation for the semantics of such conditional assessments that are arguably ascriptions of certain *conditional deontic statuses*.[126] With this, I have filled the toolbox with the tools I need for this part. We can turn now to a deeper and better analysis of the CHALLENGE in all its facets.

[126] In our everyday language there are certainly different kinds of "Iffy Oughts", as Feldman (1980, ch. 4) called them. The ones introduced here are the kind relevant to the CHALLENGE, and I do not enter the field any further (but I claim that they bear an important structural resemblance to the conditional oughts Jeff Horty explored in chapter 5 of Horty 2001).

Chapter 6

The Real Challenge

I begin the development of my own approach by dismissing the Internal Challenge. For this, we revisit the Argument and give reasons for why it is invalid.

As we will see, this result offers much less help to proponents of moac than one might initially expect, as it gives rise to a new challenge: the Real Challenge, which lends its name to this chapter. This challenge is more fundamental and can be understood as the *cause* of the Challenge, rendering the latter more of a mere *symptom*.

The general insight is that, due to the apparent inseparability of Troublemakers, consequentialists have opted for the argumentative strategy discussed in detail in Subsubsection 3.5.2.3: incorporating facts about what other agents do, will do, or have done into the actual context of these collective decision situations. I shall refer to this as the Consequentialist Standard Move (or, more concisely, the csm).

The essential point of this chapter, within the context of my overall project, then, is as follows: to avoid a fundamental challenge, consequentialism took a wrong turn—the csm—which ultimately led to the emergence of the Challenge in the first place. Yet, without the csm, camp moac must face a myriad of systematic deontic gaps, effectively accepting an untenable degree of deontic *in*completeness. This is the Real Challenge, and it is why proving the Argument invalid in this sense is indeed a Pyrrhic victory for my consequentialist project. The remaining challenge, therefore, is to resolve the Real Challenge without inadvertently recreating the Challenge.

Before addressing this challenge, I would like to briefly present one further observation highlighting why the current treatment of collective decision situations, i.e., the csm, seems flawed. This serves as an independent motivation to fundamentally question the csm, separate from concerns about the Challenge and, more specifically, from mh.

6.1 The Principle of Moral Balance

PMH is not the only principle threatened by collective contexts that proponents of MOAC arguably should care about. Another candidate, one I have not been able to find in the literature, is based on the intuition that if two combinations of actions lead to morally equivalent consequences, the moral assessments of the individual actions making up these combinations must not 'diverge too much'.

I deliberately leave open what exactly constitutes a divergence that is 'too large', as I wish to avoid committing to a specific interpretation here. However, this vagueness can make it challenging to articulate the general idea clearly. To provide a clearer and more accessible explanation, consider the most extreme form of divergence and thus the weakest form of PMB (which suffices for the purposes of this section): when all actions in one combination are deemed right and all actions in the other combination are deemed wrong. This severe imbalance represents the maximal divergence.

Call this general idea—which constrains the moral assessments of actions that, in combination, lead to morally equivalent outcomes—the Principle of Moral Balance (or PMB short). Without PMB being true in some sense, the connection between the quality of consequences and the moral status of actions, a cornerstone of consequentialist thought, is undermined. Here is my suggestion for a concrete and concise formulation, grounded in the most extreme form of divergence:

Principle 6.1 (Moral Balance (MB)) *Let D be a collective decision situation, and let Υ and Υ' be two proper combinations in D. If Υ and Υ' lead to morally equivalent consequences, then it cannot be that all $\Phi \in \Upsilon$ are (necessarily) right and all $\Phi' \in \Upsilon'$ are (necessarily) wrong.*

We have already seen a case where MB is apparently violated. Recall Glover's two Beans and Bandits cases (see page 64). The first one is about bandits who obviously have never heard of the Challenge (or just don't care if what they are doing is wrong or not, or who are not from camp MOAC …), recall:

Case 3.4 (Beans and Bandits– One to One) *Imagine a village with 100 very hungry, nearly starving tribesmen, each preparing their lunch on one of 100 small fireplaces. 100 mildly hungry bandits are waiting outside the village for the right moment to steal the villagers' food. While the villagers are briefly distracted and turn their backs on their fireplaces, the bandits sneak into the village unnoticed. Each thief steals one villager's bowl to satisfy their appetite.*

There was no particular challenge in condemning these bandits' actions with MOCOR. Now recall the case where the bandits behave rather strangely (per-

haps they have been tutored by a wandering practical philosopher in the meantime):

Case 3.5 (Beans and Bandits– Many to Many) *Imagine a village with 100 very hungry, nearly starving tribesmen, each preparing their lunch on one of 100 small fireplaces. 100 mildly hungry bandits are waiting outside the village for the right moment to steal the villagers' food. While the villagers are briefly distracted and turn their backs on their fireplaces, the bandits sneak into the village unnoticed. Each thief steals one bean from each villager's bowl to satisfy their appetite.*

Regarding PMB, consider the following: Both combinations of actions result in morally equivalent outcomes—100 hungry villagers without lunch and 100 well-fed bandits. However, in the one-to-one case, MOCOR deems that each bandit has done something wrong (specifically, severely harming the corresponding villager by robbing them of their lunch), whereas in the many-to-many case, MOCOR appears to find that no one has done anything wrong. This discrepancy violates MB. Importantly, this challenge is independent of PMH, as it is not rooted in the fact that morally suboptimal results are produced in any way. Instead, it lies in the *divergence* between the respective moral assessments.

Beans and Bandits is a Cumulative Effects Case, and I have excluded these cases from this project (Section 4.2). But we can build Coordination Cases with similar structure:

Case 6.1 (Scholars' Birthday Standoff) *Markus and Caro, two prominent figures in their scientific field, share the same birthday. On this particular day, they find themselves together at a prestigious scientific conference addressing a significant and societally relevant topic. For years, Markus and Caro have been at the center of a scholarly divide, each leading a faction of dedicated followers. The researchers at the conference are evenly split in their allegiance to either Markus or Caro, and tensions between the two camps are noticeable.*

Markus and Caro each have three possible choices regarding their interactions: they can rise above their differences and wish the other a happy birthday, they can completely ignore the topic of their shared birthday, or they can openly display animosity toward one another.

From the perspective of fostering harmony at the conference, mutual birthday wishes would create the most positive atmosphere. In fact, such a gesture would lead to a cooperative spirit and even spark scientific breakthroughs capable of addressing pressing global challenges. If only one extends birthday wishes while the other remains silent, the ambiance would still be reasonably reasonably friendly, with modest progress achieved.

If both choose to ignore the birthday topic or display hostility, the conference mood would become strained, though not unexpectedly, as such behavior aligns

with the participants' expectations. While this would result in a missed opportunity, the discord would likely subside in the medium term. However, a one-sided display of hostility would have the most damaging effect, significantly souring the atmosphere and leaving a lasting negative impact. It would demoralize participants for years to come and hinder progress on critical societal issues.

Scholars' Birthday Standoff is best conceived of as a Coordination Case represented by the following normal form:

		Markus		
		congratulate	say nothing	be rude
Caro	congratulate	+++	+	−−
	say nothing	+	−	−−
	be rude	−−	−−	−

There are two combinations of actions with morally equivalent outcomes in this case. Whether Markus and Caro both avoid the topic of their shared birthday and remain silent, or whether they both act rudely as expected, the outcomes are morally equivalent. However, applying the CSM—adding what the one agent does, did, or will do to the context for assessing the other agent's actions—yields contrasting moral assessments under MOCOR: In the first case (where both remain silent), MOCOR judges their actions as wrong because Markus (and symmetrically, Caro) could have improved the situation simply by congratulating the other. In contrast, in the second case (where both act rudely), MOCOR deems their actions right. This is because, given the mutual rudeness, Markus (and likewise Caro) could not have made a positive difference by acting differently. In fact, any unilateral deviation—whether congratulating or remaining silent—would have escalated tensions further, spoiling the conference atmosphere entirely. Thus, while the CSM generates assessments consistent with its internal logic, these judgments violate PMB, as they assign drastically different moral statuses to actions that, in combination, lead to morally equivalent outcomes. Once again, the CSM produces results unacceptable for camp MOAC, this time due to a violation of PMB rather than PMH.

I do not want to go deeper into PMB or MB here. Instead, they should motivate us to question precisely this kind of (anticipatory) retrospective reasoning, i.e., the CSM, independent from the Challenge itself. With this doubt in mind, we return to the Challenge—and more specifically to the Internal Challenge and the Argument, which rely heavily on this reasoning.

6.2 Revisiting the ARGUMENT

Recall, once again,

The ARGUMENT – tentative

$P_{\exists\text{TROUBLE}}$: There are Troublemakers: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

P_{MOCOR}: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

P_{MH}: If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (according to this theory).

$C_{\neg\text{ADEQ}}$: MOAC is not an adequate moral theory.

As announced, my goal is to cast doubt on the deductive validity of the ARGUMENT. An argument is (deductively) valid if and only if the truth of its premises (deductively) entails the truth of its conclusion. In other words, it cannot be that the premises of a valid argument are true but its conclusion is false. Thus, to decide whether the ARGUMENT is valid, we must determine whether the conclusion follows (deductively) from its premises. Because the validity of an argument is a structural, semantic property of an argument, we should take a careful look at its semantic structure. We are thus searching for an adequate formal representation, which typically requires careful interpretation for complex arguments. Thankfully, we have already worked to understand the ARGUMENT (Section 3.5) and developed quite some formal toolbox (Chapter 5). However, before I turn to the formal representation of the ARGUMENT, I will highlight some essential but usually overlooked details and sketch the general intuition behind the interpretation I later advocate.

6.3 The Intuition: Gaps Filled Badly

Before I establish the invalidity of the ARGUMENT based on the formal toolbox introduced so far, I will briefly convey the general insight on which my

reconstruction is based. First, we can assume, for now, the existence of Troublemaker. Suppose the agents in some Troublemakers realize a troublesome combination. Accordingly, they produce a suboptimal outcome (Collective Suboptimality), and, given the actions of the other agents, none of the agents involved could have made a difference for the better by unilaterally acting differently (Individual Optimality). The Challenge is based on two observations: on the one hand, in light of Individual Optimality, mocor seems to yield that all agents have acted rightly; on the other hand, mh requires at least one wrong action in light of Collective Suboptimality.

Second, we have established that the Internal Challenge is grounded in the two principles mocor and mh. If both mocor and mh are true, the matter seems clear: moac is inadequate. This might tempt one to devise a rough-and-ready formalization of the Argument, such as the following:[127]

The Argument (formally, naïve)

$$P_{\exists\text{trouble}}\colon\quad \exists\,\mathsf{D} = \langle \mathcal{A},\, \Gamma,\, \mathsf{Out}_C :\, \Psi \,\to \mathcal{O}\rangle \,:\, \exists\,\Upsilon \in \Psi :\; \neg\text{GlobOpt}(\Upsilon) \wedge \text{IndiOpt}(\Upsilon).$$

$$P_{\text{mocor}}\colon\quad \forall\,\mathsf{D} = \langle \mathcal{A},\, \Gamma,\, \mathsf{Out}_C :\, \Psi \,\to \mathcal{O}\rangle \,:\, \forall\,\Upsilon \in \Psi :\; \text{IndiOpt}(\Upsilon) \to \text{All-R}(\Upsilon, \text{moac}).$$

$$P_{\text{mh}}\colon\quad \forall\, T :\, \text{Adeq}(T) \to \Big(\forall\,\mathsf{D} = \langle \mathcal{A},\, \Gamma,\, \mathsf{Out}_C :\, \Psi \,\to \mathcal{O}\rangle \,:\; \big(\forall\,\Upsilon \in \Psi :\, \neg\text{GlobOpt}(\Upsilon) \to \neg\text{All-R}(\Upsilon, T)\big)\Big).$$

$$C_{\neg\text{adeq}}\colon\quad \neg_{\text{adeq}}(\text{moac})$$

If this were an adequate formal representation of the Argument, validity would be beyond question. $P_{\exists\text{trouble}}$ together with P_{mocor} entails the existence of collective decision situations where a suboptimal combination of actions consists only of right actions, i.e.,

$$\exists\,\mathsf{D} = \langle \mathcal{A},\, \Gamma,\, \mathsf{Out}_C :\, \Psi \,\to \mathcal{O}\rangle \,:\, \exists\,\Upsilon \in \Psi :\; \neg\text{GlobOpt}(\Upsilon) \wedge \text{All-R}(\Upsilon, \text{moac}).$$

[127] Here, the predicates are defined as follows: GlobOpt(…) denotes that a combination of actions produces the best possible consequences (and thus ¬GlobOpt(…) corresponds to Collective Suboptimality); IndiOpt(…) corresponds to Individual Optimality; All-R(…,…) indicates that all actions within a combination are deemed right according to a given theory; and Adeq(…) signifies that a theory is adequate.

By exemplifying P_{MH} with MOCOR, we can then infer:

$$\text{Adeq}(\text{MOAC}) \rightarrow \exists\, \mathsf{D} = \langle \mathcal{A},\, \Gamma,\, \mathsf{Out}_C :\, \Psi \,\rightarrow \mathcal{O} \rangle \,:\, \exists\, \Upsilon \in \Psi : \\ \neg\text{All-R}(\Upsilon, \text{MOAC}) \wedge \text{All-R}(\Upsilon, \text{MOAC}).$$

Since the right side of the conditional is a logical falsehood—there can be no collective decision situation with a combination in which *all* actions are right and *not* all actions are right at the same time— we get the result that MOAC is inadequate.

However, at a second glance and against the background of what we learned about the structure of collective decision situations in Chapter 5, the above reconstruction is overly simplistic. Most importantly, MOCOR only arrives at the alleged assessments if, in assessing each action, it is already considered as a matter of fact what all other agents have done. The assessment is made in *retrospect*. This is what Consequentialist Standard Move (CSM) was all about: adding facts about what the other agents do, will do, or have done to the actual context in order to 'escape' the conditional.

MH, however, relies on the plausibility of Collectively Maximizing. Recall:

Principle 5.1 (Collectively Maximizing) *Let D be a collective decision situation with domain Ψ and with actual context C. If $\Upsilon \in \Psi$ consists only of right actions, then there is (and can be) no alternative $\Upsilon' \in \Psi$ with better consequences than Υ relative to C.*

Collectively Maximizing arguably has a kind of guiding, anticipatory, prospective character. Accordingly, in line with all that we have learned in the first part of this thesis, MH is meant to express the idea that morality has the function of implicitly *coordinating* our behavior in such a way as to produce the best outcome we can collectively produce. A few witnesses shall be called again. First, here is Fred Feldman (1980, p. 168):

> Quite a few moral philosophers seem to believe that when all the members of a social group do what they morally ought to do, the group as a whole does benefit more than it would have from the performance of any worse alternative set of actions. I shall say that any such view is a version of the Principle of Moral Harmony (PMH).

Similarly, but much more recently, Douglas Portmore (2018, p. 12) reads MH like this:

> The principle of moral harmony holds that a moral theory must be morally harmonious—that is, it must be such that the agents who satisfy the theory, whoever and however numerous they may be, are guaranteed to produce the morally best world that they together have the option of producing.

Or remember Stephen Toulmin's dictum (1953, p.137) according to which "we can provisionally define [the 'function' of ethics] as being 'to correlate our feelings and behavior in such a way as to make the fulfillment of everyone's aims and desires as far as possible compatible'." It stands to reason that Toulmin here means that our behavior should be aligned, coordinated, attuned, …, rather than 'retroactively assessed'.

Thus, MH is first of all about where we arrive when we (with or without the intention to do so) *follow* morality, i.e., what we bring about when everything we do *satisfies* the moral theory under consideration (to use Regan's, Parfit's, and Portmore's lingo), if we select from the *set of right actions* (Regan's lingo again), if 'doing what one ought' (Feldman) *according* to a theory, etc., etc. *This is not a backward-looking perspective.*

Therefore, to determine whether there are genuine violations of MH in Troublemakers, we should proceed in the same manner that Regan *attempted* to apply in the context of his impossibility result: We should examine each set of right actions for each agent as implied by MOAC. If we find a combination of actions that consists solely of right actions selected in that manner, then it follows that if the agents were to act accordingly, they would be doing what they ought to do (in the sense of Feldman's quotation). Consequently, they would be *fulfilling* MOAC (in the sense of Portmore's and Regan's formulations). Now, if this combination of actions were also to result in suboptimal outcomes, we could confidently conclude that Collectively Maximizing has been violated—and therefore, by MH, that MOAC would be inadequate.

For troublesome combinations to be *truly* troublesome, thus, they must consist solely of right actions for the agents within a Troublemaker—full stop. This means that no additional facts (in the sense of boundary conditions) about the actions of other agents are introduced. Put differently, for the Argument to hold, there must exist right actions for the agents, i.e., Troublemakers must be resolvable.

But can these forward-looking assessments truly be warranted by or inferred from the backward-looking reasoning in P_{MOCOR}? My claim is that they cannot. To apply MOCOR, we must identify individual decision situations for all agents. As demonstrated in the previous chapter, this can indeed be achieved by reducing the collective decision situation through the addition of other agents' actions to the context. However, as we have seen, the assessments derived using this technique of conditionalizing only warrant conditional assessments relative to the original collective decision situation. What we actually need is to identify right actions for all agents in a Troublemaker *without* relying on the CSM.

As a result, even if, in retrospect and within the context where all agents have already acted, every action *was* right, this does not at all imply that any of the agents' actions *must* have been right at the relevant decision time.

Consequently, the assessments referenced in P_{MOCOR} are derived relative to a different context than the assessments required by MH. In this regard, the seemingly trivial inference described above does not do justice to the ARGUMENT.

Before I turn to a better reconstruction of the ARGUMENT, let me emphasize that the actions of the other agents are not already part of the actual context. Assuming that they are is to assume determinism. But assuming determinism violates

Principle 6.2 (Methodological Indeterminism) *The question of what is right to do for an agent in a decision situation is pointless if what the agent does is already predetermined. Even if determinism were right, in the context of morality, we should* pretend *that it is not.*

I will not do much to defend this principle because I consider it to be quite self-evident. However, one quick justification for Methodological Indeterminism should be mentioned: We generally presuppose the falsity of determinism in our moral discourse. More specifically, we assume that, in any given decision situation, the agents involved have the ability to perform any of their available options. This assumption is particularly central to consequentialist reasoning, which relies heavily on comparing the possible consequences of one option with those of another. However, when assessing the options available to all agents within a collective decision situation, we face a dilemma: to evaluate one agent's options, we must assume the actions of the other agents as fixed, and vice versa. This leads to a situation where we effectively assume the actions of *all* agents, leaving no genuine choice left to assess. I claim that this provides a sufficiently strong reason for proponents of MOAC to endorse Methodological Indeterminism.

Naturally, I am not the first one to notice this struggle. Here is a passage[128]

[128]The context of Prior's quote is also interesting concerning Methodological Indeterminism. Prior starts from a point Moore made in his *Principia Ethica* (Moore 1903) concerning determinism and its importance for consequentialist considerations:

> My argument is dilemmatic. Either determinism is true or it is not. If determinism is true then there are not really (though there may seem to be) a number of alternative actions which we could perform on a given occasion; the one action that we can perform is the one that we do perform. Hence whatever we in fact do is the best possible action (the one with the best possible total consequences) because it is the *only* possible action; so that whatever we in fact do is our duty, in Moore's sense of 'duty.' Moore himself saw this horn of the dilemma (and indeed it is a commonplace that determinism presents problems of this sort); but it has another horn which so far as I know he did *not* see. [..., the passage below]
>
> The conclusion seems clear. If determinism is true, then whatever we do is our duty in Moore's sense of "duty," and if determinism is not true then nothing at all is our duty in this sense.

from A. N. Prior (1956, pp. 91–92; see also Horty, 2001, chap. 4):

> Suppose that determinism is *not* true. Then there may indeed be a number of alternative actions which we could perform on a given occasion, but none of these actions can be said to have any "total consequences," or to bring about a definite state of the world which is better than any other that might be brought about by other choices. For we may presume that other agents are free beside the one who is on the given occasion deciding what he ought to do, and the total future state of the world depends on how these others choose as well as on how the given person chooses; and even if there were not other people to spoil one's calculations there would still be oneself, with one's own future choices, or some of them, undetermined like this present one (unless a man decides that it is too risky for him to have any further freewill, and on this very ground finds it to be his duty to do away with himself). And while I speak here of one's calculations being spoilt, the trouble of course goes deeper than that—it's not merely that one cannot calculate the totality of what will happen if one decides in a certain way; the point is rather that there *is* no such totality.
>
> The conclusion seems clear. If [...] determinism is not true then nothing at all is our duty in this sense.

Call this challenge the Real Challenge. It is not that moac would make *incorrect* assessments in Troublemakers, which then violates the ideal of pmh, but rather that moac simply does not make *any assessments at all* in these cases. This, of course, is also not in accordance with the spirit of pmh, but does not spawn any violations, though. But worse, moac, instead of having Deontic Completeness, is a Swiss cheese, full of systematic deontic voids, gaps, and holes.

All this might be a bit abstract, so we return to our running example, Two Factories, before turning to another reconstruction attempt. Recall the normal form:

		Ben	
		pollute	produce cleanly
Ann	pollute	second-worst	worst
	produce cleanly	worst	best

We now have a sufficient understanding of the situation to be sure that we can indeed derive all of the following propositions from Two Factories:

The solution proposed in this part of my dissertation can thus be understood as a resolution of Prior's dilemma. It allows champions of moac to adopt Methodological Indeterminism, i.e., to work under the assumption of the falsity of determinism, without running into the problem raised here by Prior, which is ultimately the Real Challenge.

(30) If Ann pollutes, it is right for Ben to pollute.

(31) If Ben pollutes, it is right for Ann to pollute.

(32) Once Ann has polluted, it will have been right for Ben to have polluted.

(33) Once Ben has polluted, it will have been right for Ann to have polluted.

(34) If Ann produces cleanly, it is right for Ben to produce cleanly.

(35) If Ben produces cleanly, it is right for Ann to produce cleanly.

(36) Once Ann has produced cleanly, it will have been right for Ben to have produced cleanly.

(37) Once Ben has produced cleanly, it will have been right for Ann to have produced cleanly.

Similarly, relative to Two Factories together with the fact that both pollute, i.e., that they instantiate the troublesome combination within Two Factories, all of the following propositions are true:

(38) That Ann polluted was right.

(39) That Ben polluted was right.

(40) That Ann polluted was right because Ben polluted.

(41) That Ben polluted was right because Ann polluted.

However, the *following* propositions are *false* relative to Two Factories together with the fact that both pollute:

(42) That Ann polluted was right, independently of what Ben did.

(43) That Ben polluted was right, independently of what Ann did.

In essence, this highlights that while MOAC can provide numerous assessments, it does *not* yield any non-conditional assessments relative to Two Factories—not even in retrospect. This remains true, I remind you, as long as Two Factories is considered without incorporating the actually performed actions. This result is a structural one and can be generalized to apply to (many options of many agents within) non-decomposable collective decision

situations, including Troublemakers. The issue, therefore, is not the Challenge, as no violation of MH can arise in the absence of resolvability. The true problem with Troublemakers lies in MOAC's inability to provide *any* non-conditional assessments in such cases to begin with. This fundamental shortcoming is what I refer to as the Real Challenge.

At this point, one might object that all this is quite 'hand-wavy'—and rightly so. I owe the reader a better, more careful, and formal reconstruction of the Argument.

6.4 The Logical Structure of The Argument

In this section, I will develop a reasonable formalization of the Argument. I turn to each of the three premises, one by one. Only then do I put them together and show the invalidity of the Argument.

6.4.1 The Structure of $P_{\exists\text{TROUBLE}}$: Straightforward

$P_{\exists\text{TROUBLE}}$'s structure is easily captured based on the semi-formalism introduced earlier. For this, first recall the tentative definition from the introduction:

Definition 1.1 (Troublemakers – tentative) *A collective decision situation is a Troublemaker if and only if there is a troublesome combination of options therein, i.e., the agents can act in ways such that*

> **(Collective Suboptimality)** *together they would produce a morally suboptimal outcome and*
>
> **(Individual Optimality)** *none of them could make a difference for the morally better by unilaterally acting differently.*

Acknowledging that being a Troublemaker depends on axiological questions allows us to be slightly more precise and more general:

Definition 6.1 (Troublemaker (relative to some axiology V)) *A collective decision situation is a Troublemaker (relative to some axiology V) if and only if there is a troublesome combination of options therein (relative to some axiology V), i.e., if the agents can act in ways such that*

Collective Suboptimality *together they would produce a morally suboptimal outcome (relative to V) and*

Individual Optimality *none of them could make a difference for the morally better (relative to V) by unilaterally acting differently.*

To reveal the structure of $P_{\exists \text{TROUBLE}}$, I need a formal definition of troublesome combinations. (While formalizing Troublemaker-hood presents an excellent opportunity to achieve a higher level of generality, this step is not strictly necessary, as it would not alter the core of my argument. Instead, as noted in the previous chapter, I will continue to restrict my focus to minimal collective decision situations. More specifically, I concentrate on Two Factories–like cases, i.e., Coordination Cases that satisfy the Triad.)

Building on the auxiliary notion of sets $\Psi^{\Upsilon,1}$ of 1-variants of a combination $\Upsilon \in \Psi$ (cf. Subsection 5.2.2), we define, given a decision situation $\mathsf{D} := \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O} \rangle$ and a valuation function $\mathsf{Val} : \mathcal{O} \to \mathcal{V}$ (with an order $<$ over $\mathcal{V}$), the *set of troublesome combinations* as

$$\Psi_{\text{trouble}} := \Big\{ \Upsilon \in \Psi \mid \overbrace{\exists\, \Upsilon' \in \Psi : \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big) > \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon\big)\big)}^{\approx \text{ Collective Suboptimality}} \wedge \underbrace{\nexists\, \Upsilon' \in \Psi^{\Upsilon,1} : \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big) > \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon\big)\big)}_{\approx \text{ Individual Optimality}} \Big\}.$$

This can be rewritten, using the arg max notion, as

$$\Psi_{\text{trouble}} := \Big\{ \Upsilon \in \Psi \mid \overbrace{\Upsilon \notin \operatorname*{arg\,max}_{\Upsilon' \in \Psi} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big)}^{\approx \text{ Collective Suboptimality}} \wedge \underbrace{\Upsilon \in \operatorname*{arg\,max}_{\Upsilon' \in \Psi^{\Upsilon,1}} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big)}_{\approx \text{ Individual Optimality}} \Big\}.$$

Accordingly, a decision situation D with $\mathsf{D} := \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O} \rangle$ is a Troublemaker if and only if $\Psi_{\text{trouble}} \neq \emptyset$, i.e., if and only if there is a troublesome combination.

Thus, we get a formal structure for $P_{\exists \text{TROUBLE}}$ that reads:

$$(P_{\exists \text{TROUBLE}}) \quad \exists\, \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O} \rangle : \exists\, \Upsilon \in \Psi : \Upsilon \notin \operatorname*{arg\,max}_{\Upsilon' \in \Psi} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big) \wedge \Upsilon \in \operatorname*{arg\,max}_{\Upsilon' \in \Psi^{\Upsilon,1}} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big)$$

Before we turn to the semantic structure of P_{MOCOR}, it is worth stressing that the formal condition for troublesome combinations illustrates that we can frame the CHALLENGE as an optimization issue, i.e., MOAC's apparent inability to guarantee optimal solutions when used as a decision procedure. This allows us to understand Troublemakers as challenging insofar as they allow for combinations that are local maxima. These local maxima are where MOCOR allegedly 'gets stuck'. To drive this point home, consider

Definition 6.2 (Local and global Maxima)
Let $\mathsf{D} := \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$ *be a collective decision situation.*

A combination of actions $\Upsilon \in \Psi_\Gamma$ *is a local (or relative) maximum combination if and only if* Υ*'s consequences are a local (or relative) maximum within the*

valuative profile of D, i.e., there is no individual deviation from that combination which would have better consequences in D.

A combination of actions is a global *(or absolute)* maximum combination *if and only if Υ's consequences are a global maximum within the evaluative profile of D, i.e., there is no combination with better consequences in D.*

Trivially, every global maximum combination is a local one: If there is no combination with better consequences, then there cannot be a combination as a result of unilateral deviation that has better consequences. A bit more formally, then, we get, for some decision situation D (as above) and a valuation function $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$, that

$$\begin{gathered}\Upsilon \in \Psi_\Gamma \text{ is a local maximum (combination)}\\ \text{if and only if}\\ \text{there is no } \Upsilon' \in \Psi^{\Upsilon,1} \text{ with } \mathsf{Val}\big(\mathsf{Out}_C\left(\Upsilon'\right)\big) > \mathsf{Val}(\mathsf{Out}_C(\Upsilon))\\ \text{if and only if}\\ \Upsilon \in \underset{\Upsilon' \in \Psi^{\Upsilon,1}}{\arg\max}\, \mathsf{Val}\big(\mathsf{Out}_C\left(\Upsilon'\right)\big)\end{gathered}$$

and

$$\begin{gathered}\Upsilon \in \Psi_\Gamma \text{ is a global maximum (combination)}\\ \text{if and only if}\\ \text{there is no } \Upsilon' \in \Psi_\Gamma \text{ with } \mathsf{Val}\big(\mathsf{Out}_C\left(\Upsilon'\right)\big) > \mathsf{Val}(\mathsf{Out}_C(\Upsilon))\\ \text{if and only if}\\ \Upsilon \in \underset{\Upsilon' \in \Psi_\Gamma}{\arg\max}\, \mathsf{Val}\big(\mathsf{Out}_C\left(\Upsilon'\right)\big).\end{gathered}$$

Troublesome combinations are precisely those combinations of actions that are local but not global maxima. Put differently, D qualifies as a Troublemaker if and only if it contains a genuine local maximum that is not a global maximum. Derek Parfit touched on a similar idea when he proposed classifying different types of Coordination Cases based on their "contour maps" (cf. ibid., pp. 14–15; see also Figure 6.1). In this context, Bacharach's name "Hi-Lo Cases" for such cases is particularly fitting (Bacharach 1999, p. 130).

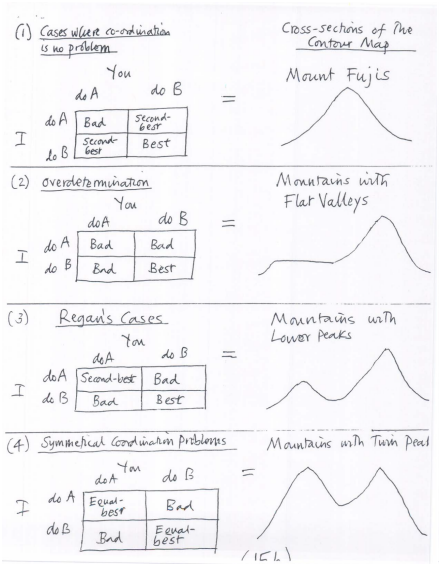


Figure 6.1: Parfit's contour maps (Parfit 1988, p. 15B (sic!)).

6.4.2 The Structure of P_{MOCOR}: Ex Post!

Next, let us consider

(P_{MOCOR}) If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

Since this proposition obviously is meant to tell us something about all decision situations (with specific properties), we can start by getting the quantifiers right:

(P_{MOCOR})
$\forall\, \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O} \rangle : \forall \Upsilon \in \Psi :$
if none of the agents in $\mathcal{A}$ could make a difference for the better by unilaterally acting differently,
then each of them would act morally right.

Up to this point, no doubt, the naïve formalization above was on the right track. However, drawing on what we learned about $P_{\exists\text{TROUBLE}}$, we can easily translate the antecedent in a syntactically richer way than it was offered then:

(P_{MOCOR})
$\forall\, \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O} \rangle : \forall \Upsilon \in \Psi :$
if $\nexists\, \Upsilon' \in \Psi^{\Upsilon,1} : \mathsf{Val}\big(\mathsf{Out}_C(\Upsilon')\big) > \mathsf{Val}\big(\mathsf{Out}_C(\Upsilon)\big)$
then each of them would act morally right.

For the formalization of the consequent, we make use of two other building blocks introduced earlier. First, we need the definition of the predicate of rightness relative to a decision situation, a context, and a moral theory (cf. page 25), i.e., $T, \mathsf{D}, C \vDash R\phi$. Second, we need to apply the apparatus we introduced to capture the technique of conditionalization defined in the preceding chapter:

(P_{MOCOR})
$$\forall\, \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O} \rangle : \forall \Upsilon \in \Psi :$$
$$\Big(\nexists\, \Upsilon' \in \Psi^{\Upsilon,1} : \mathsf{Val}\big(\mathsf{Out}_C(\Upsilon')\big) > \mathsf{Val}\big(\mathsf{Out}_C(\Upsilon)\big) \to \forall \phi \in \Upsilon : \text{MOAC}, D_{\downarrow \Upsilon \ominus \phi}, [\![C \oplus (\Upsilon \ominus \phi)]\!] \vDash R\phi \Big).$$

Alternatively, we can express this in terms of local maxima:

(P_{MOCOR})
$$\forall\, \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O} \rangle : \forall \Upsilon \in \Psi :$$
$$\Big(\Upsilon \in \operatorname*{arg\,max}_{\Upsilon' \in \Psi^{\Upsilon,1}} \mathsf{Val}\big(\mathsf{Out}_C(\Upsilon')\big) \to \forall \phi \in \Upsilon : \text{MOAC}, D_{\downarrow \Upsilon \ominus \phi}, [\![C \oplus (\Upsilon \ominus \phi)]\!] \vDash R\phi \Big).$$

The underlying idea is essentially the CSM: each agent's action is assessed within the context of all other agents' actions, which collectively form a combination where no unilateral deviation could lead to a better outcome. This enrichment of the original context effectively reduces the collective decision situation to individual decision situations, which can then be evaluated by MOAC. Since the antecedent ensures that no unilateral deviation can make a positive difference, the action of the agent in question is consequently assessed as right according to MOAC.

6.4.3 The Structure of P_{MH}: Ex Ante

We can now address the final remaining premise:

(P_{MH}) If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (according to this theory).

Recall the more precise notion of Collectively Maximizing we introduced in the last chapter:

Principle 5.1 (Collectively Maximizing) *Let D be a collective decision situation with domain Ψ and with actual context C. If $\Upsilon \in \Psi$ consists only of right actions, then there is (and can be) no alternative $\Upsilon' \in \Psi$ with better consequences than Υ relative to C.*

Building on Collectively Maximizing and the considerations outlined in Section 6.3, along with the formalizations introduced earlier—most notably the formalization of P_{MOCOR}—we can proceed directly to the formalization of MH. First, let us recall the principle of MH itself:

Criterion 1.2 (Moral Harmony (MH) – tentative) *A moral theory is adequate only if it is true that if all agents act rightly (according to this theory), then they are guaranteed to produce the morally best outcome they could bring about together.*

The straightforward formalization of MH then looks like this:

$$(P^{\leftarrow}_{\text{MH}}) \qquad \begin{aligned} &\forall T : \big(\text{Adeq}\, T \rightarrow \forall \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \rightarrow \mathcal{O}\rangle : \forall \Upsilon \in \Psi : \\ &\quad \forall \phi \in \Upsilon : T, D, C \vDash R\phi \rightarrow \Upsilon \in \underset{\Upsilon' \in \Psi}{\arg\max}\, \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big)\big). \end{aligned}$$

Now, to get a proper formalization of P_{MH}, we only need to take the contraposition of the inner conditional. This yields (presupposing the shallow wrongness predicate from Subsection 2.3.4):

$$(P_{\text{MH}}) \qquad \begin{aligned} &\forall T : \big(\text{Adeq}\, T \to \forall \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O}\rangle : \forall \Upsilon \in \Psi : \\ &\quad \Upsilon \notin \operatorname*{arg\,max}_{\Upsilon' \in \Psi} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big) \to \exists \phi \in \Upsilon : T, D, C \models W\phi\big). \end{aligned}$$

Note that this formalization does not involve any conditionalizing on further facts. Most importantly, we do *not* have

$$(P^{\to}_{\text{MH}}) \qquad \begin{aligned} &\forall T : \big(\text{Adeq}\, T \to \forall \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O}\rangle : \forall \Upsilon \in \Psi : \\ &\quad \Upsilon \notin \operatorname*{arg\,max}_{\Upsilon' \in \Psi} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big) \\ &\qquad \to \exists \phi \in \Upsilon : T, D_{\downarrow \Upsilon \ominus \phi}, [\![C \oplus (\Upsilon \ominus \phi)]\!] \models W\phi\big). \end{aligned}$$

Now that we have plausible and defensible formalizations of the three premises, we can turn to the overall ARGUMENT.

6.4.4 Putting Things Together

If we put the three formalizations together (and add the very trivial conclusion), we get a reasonable formalization of the whole ARGUMENT:

The ARGUMENT (formally)

$P_{\exists\text{TROUBLE}}$:
$$\begin{aligned} &\exists \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O}\rangle : \exists \Upsilon \in \Psi : \\ &\quad \Upsilon \notin \operatorname*{arg\,max}_{\Upsilon' \in \Psi} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big) \wedge \Upsilon \in \operatorname*{arg\,max}_{\Upsilon' \in \Psi^{\Upsilon,1}} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big) \end{aligned}$$

P_{MOCOR}:
$$\begin{aligned} &\forall \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O}\rangle : \forall \Upsilon \in \Psi : \\ &\quad \Big(\Upsilon \in \operatorname*{arg\,max}_{\Upsilon' \in \Psi^{\Upsilon,1}} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big) \\ &\qquad \to \forall \phi \in \Upsilon : \text{MOAC}, D_{\downarrow \Upsilon \ominus \phi}, [\![C \oplus (\Upsilon \ominus \phi)]\!] \models R\phi\Big). \end{aligned}$$

P_{MH}:
$$\begin{aligned} &\forall T : \big(\text{Adeq}\, T \to \forall \mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O}\rangle : \forall \Upsilon \in \Psi : \\ &\quad \Upsilon \notin \operatorname*{arg\,max}_{\Upsilon' \in \Psi} \mathsf{Val}\big(\mathsf{Out}_C\big(\Upsilon'\big)\big) \to \exists \phi \in \Upsilon : T, D, C \models W\phi\big). \end{aligned}$$

$C_{\neg\text{ADEQ}}$: $\neg$Adeq MOAC

It is evident that the conclusion cannot be logically derived from the premises (though we will examine this in greater detail in a second). The ARGUMENT, therefore, is invalid. However, somewhat surprisingly, this does not signify

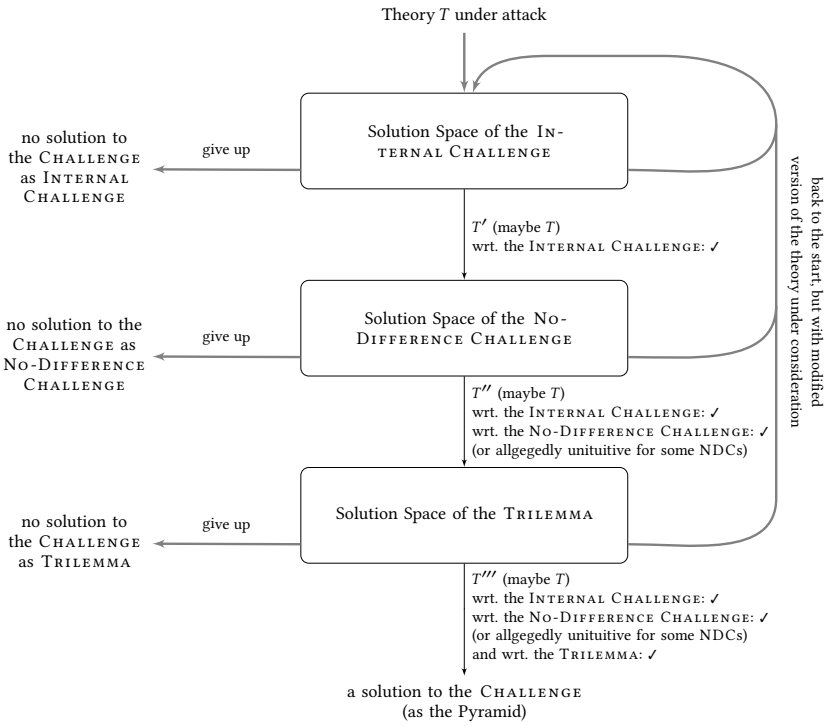


Figure 6.2: The solution space of the Challenge as the Pyramid.

the resolution of the Challenge, particularly in its internal form, Internal Challenge.

This sobering assertion requires further explanation. Formally, Internal Challenge appears to be resolved, and we could, in principle, move on to addressing the No-Difference Challenge (as outlined in Figure 6.2 from Chapter 4). However, upon closer inspection, the matter proves far more complicated, as this apparent victory still leaves a significant burden on moac: it must be demonstrated that in every Troublemaker, if a troublesome combination were realized, at least one action within it would not be right.

To understand why this requires further demonstration, consider the following inference that we can draw from above reconstruction:[129] Thus, from $P_{\exists\text{TROUBLE}}$ and P_{MH}, we *can* infer that:

[129]The reasoning follows this structure: An instantiation of P_{MH} with moac can be rewritten into a formula of the form $\neg Fa \vee \forall x : (Gx \rightarrow Hx)$. This can be further reformulated as $\forall x : \neg Fa \vee (Gx \rightarrow Hx)$, which itself can be rewritten to $\forall x : (Fa \wedge Gx) \rightarrow Hx$. Meanwhile, $P_{\exists\text{TROUBLE}}$ can be expressed as $\exists x : Gx \wedge Ix$. Combining these, we infer $\exists x : (Fa \rightarrow Hx) \wedge Gx \wedge Ix$.

$$
\begin{aligned}
&\exists\, \mathsf{D} := \langle \mathcal{A},\, \Gamma,\, \mathsf{Out}_C :\, \Psi \,\to \mathcal{O}\rangle \,:\, \exists\, \Upsilon \in \Psi :\\
\text{(Gaps!)}\qquad &\quad (\text{Adeq } \textsc{moac} \to \exists\, \phi \in \Upsilon :\, \textsc{moac}, D,\, C \,\vDash\, W\, \phi)\\
&\quad \wedge\, \Upsilon \notin \underset{\Upsilon' \in \Psi}{\arg\max}\, \mathsf{Val}\big(\mathsf{Out}_C\,(\Upsilon')\big) \wedge \Upsilon \in \underset{\Upsilon' \in \Psi^{\Upsilon,1}}{\arg\max}\, \mathsf{Val}\big(\mathsf{Out}_C\,(\Upsilon')\big)
\end{aligned}
$$

Translated back into ordinary language, this means that there are Troublemakers where MOAC, in order to avoid inadequacy according to its own standards, must identify wrongdoing in those situations.

In the remainder of this chapter, I will argue that MOAC actually cannot fulfill this condition and that MOAC's structural inability to fulfill (Gaps!) substantially leads the way to a new collective challenge: The Real Challenge.

6.5 The Villain Finally Enters the Stage: The Real Challenge

At this point, we have established that MOAC does *not* assess all actions within a troublesome combination as right when that combination is performed. In this sense, there is no direct violation of MH. However, MOAC fails to satisfy a critical requirement implied by MH. The challenge that remains, then, is to demonstrate that at least one agent acts wrongly when a troublesome combination of actions is performed within a Troublemaker. It is worth recalling that MH was initially introduced in a positive formulation, as follows:

Criterion 1.2 (Moral Harmony (MH) – tentative) *A moral theory is adequate only if it is true that if all agents act rightly (according to this theory), then they are guaranteed to produce the morally best outcome they could bring about together.*

However, P_{MH} is using MH in a negative version with a contrapositive of the original condition:

Criterion 1.3 (Moral Harmony (MH) – tentative, contraposition) *If a moral theory is adequate, then, if the agents in a collective decision situation produce a morally suboptimal outcome, (necessarily) at least one of the agents acted wrongly (according to this theory).*

But as indicated already in the introduction,[130] there is a dubious translational element in this paraphrase. Assume that, instead, we were to go with

[130] Cf. footnotes 8 and 79.

Criterion 6.1 (Moral Harmony (MH, tent., contrapos., alternative)) *If a moral theory is adequate, then if the agents in a collective decision situation were to act in ways such that together they would produce a morally suboptimal outcome, then (necessarily) at least one of the agents acted* not rightly *(according to this theory).*

If we had used *this* formulation as the basis for the P_{MH} and thus the Argument, we would have ended up not with (Gaps!) above, but with

$$\begin{aligned} &\exists\, \mathsf{D} = \langle \mathcal{A},\, \Gamma,\, \mathsf{Out}_C :\, \Psi \,\to \mathcal{O} \rangle :\, \exists\, \Upsilon \in \Psi : \\ \text{(Gaps?)} \qquad &\quad (\text{Adeq MOAC} \to \exists\, \phi \in \Upsilon :\, \text{MOAC}, D,\, C \,\models\, R\, \phi) \\ &\quad \wedge\, \Upsilon \notin \underset{\Upsilon' \in \Psi}{\arg\max}\, \mathsf{Val}\big(\mathsf{Out}_C\,(\Upsilon')\big) \wedge \Upsilon \in \underset{\Upsilon' \in \Psi^{\Upsilon,1}}{\arg\max}\, \mathsf{Val}\big(\mathsf{Out}_C\,(\Upsilon')\big) \end{aligned}$$

(Gaps?) would actually be trivially fulfilled, given that one of the results of this investigation is that there is no right action whatsoever for any agent in such cases (according to MOCOR). (It should be noted that these considerations are the reason we considered the two wrongness predicates for MOAC at the end of chapter Subsection 2.3.4 (cf. page 39). A version of (Gaps!) based on the shallow consequentialist wrongness predicate W_{s} would immediately collapse into (Gaps?). At this point, only the deep consequentialist wrongness predicate W_{d} allows us to semantically distinguish the two meanings we are trying to capture.) If there is no unconditionally right action for any of the agents, obviously no agent performs a right action in a Troublemaker—whatever they do, their actions cannot be deemed right.

There are at least three reasons why the champions of MOAC should accept the stronger and thus threatening reading and thus embrace (Gaps!) instead of (Gaps). The first reason is strategic, the second is reconstructive, and the third is systematic in character.

The strategic argument is that rejecting the stronger reading is blatantly dangerous. Even if one finds no immediate justification for the stronger reading (though two such reasons will follow shortly), one risks confronting a far more threatening challenge unless a compelling case can be made for the weaker reading (and I know of none). In other words, even if proponents were to turn their back on the Internal Challenge now, they should remain uneasy. The moment someone presents a robust argument for the stronger reading, they will be forced to face that challenge regardless. Ignoring this leaves a critical vulnerability exposed.

The reconstructive argument is that the stronger reading actually corresponds to some formulations in the literature and seems to do more justice to the general idea of PMH. With respect to the first point, we can refer primarily to Pinkert's On-the-hook (cf. footnote 79 and more generally page 86), we recall:

Principle 3.8 (On-the-hook) *In any collection of agents who together gratuitously fail to bring about collectively optimal outcomes, there must be some relevant morally objectionable facts about some of the agents.*

Pinkert's formulation is explicitly meant to be a contraposition of a more general principle (Pinkert 2015, p. 976, my italics):

> I contend that for this reason, On-the-hook should not be understood as a specifically Consequentialist position. Instead, it should be understood as the *contraposition of a second-order claim about morality in general* and, hence, as a desideratum for any moral principle.

Clearly, the call for the existence of "morally objectionable facts" is closer to the idea of looking for wrongdoings. If Pinkert advocated for a version that calls for the absence of "morally commendable facts", this would correspond to Criterion 6.1's formulation of missing rightdoings.

I do not wish to rely solely on Pinkert's authority here; rather, I want to highlight that his formulation simply appears entirely correct. The core idea of PMH is inherently forward-looking.[131] PMH emphasizes that morality's role is to guide agents toward achieving the optimal outcome. Importantly, it is, in other words, not merely about *not steering toward suboptimality* but about *actively steering toward the optimal.*

Thirdly, there are systematic considerations that should let the consequentialists search for a solution which allows them to give non-conditional assessments for the agents' choices within Troublemakers. After all, a closer look at (Gaps!) reveals a deeper problem that doesn't have much to do with Troublemakers at all. Rather, the challenge ultimately lies in the fact that MOAC simply cannot provide a satisfactory answer for many overtly morally charged decisions that agents may face. MOAC fails to provide non-conditional assessments for all non-decomposable collective decision situations. This is too little for an objective moral theory that would naturally long to have Deontic Completeness, not in a purely formal but a substantive way. By this, I mean that MOAC should not retreat to the Genuine Kind View and then pretend that a property like Deontic Completeness (or even Weak Deontic Completeness) is formally attributable to MOAC because the situations of the agents in such non-decomposable collective decision situations are not individual decision situations at all. After all, these situations are such that agents must choose from a set of options. To withdraw in such situations to the position that one can only give 'it depends on what the others do'-answers, possibly

[131] This aligns with Feldman's and Portmore's citations in Section 6.3 and also resonates with Broad's positive formulation of the False Universalization principle in Subsection 3.5.1, which was designed to provide agents with a forward-looking guide for making the right decisions.

with the (lame!) excuse that the agents are not in a *real* (i.e., individual) decision situation at all, seems not only theoretically unsatisfying but comes close to committing the true Scotsman fallacy.

Thus, if we follow these reasons, we are ultimately dealing with a weakness of MOAC that extends far beyond Troublemakers and that is not directly connected to PMH even at its core. In light of Methodological Indeterminism, MOAC is simply incapable of working out which actions are right and wrong in multi-agent situations where the outcomes of the agents' actions are sufficiently interdependent. It just cannot identify *the* consequences of the individual agent's actions. MOAC can only assess retrospectively or under the assumption of already determined future actions. Both ways lead to failure.

Since we assume Methodological Indeterminism, we can conclude: MOAC fails to live up to its desideratum of Deontic Completeness (cf. Subsection 4.3.2). Instead, a closer look reveals a world of choices that takes place in moral gaps. Only when actions have filled these Gaps can it be said from the perspective of MOAC whether these actions were right or wrong. This challenge, the Real Challenge, I would like to suggest can be grasped argumentatively as follows in terms of the Real Argument:

The Real Argument

$P_{\exists \text{NON-DEC}}$:	There are non-decomposable collective decision situations, i.e., collective decision situations in which, relative to the actual and normatively sufficiently complete context, it is undefined for at least one agent, with respect to at least two of their options, how these options are to be ranked based on the moral quality of their actual consequences.
$P_{\exists \text{GAPS}}$:	If there are such collective decision situations, then there is a broad class of decision situations for which MOAC offers no meaningful guidance; in other words, it results in (a myriad of) systematic deontic gaps.
$P_{\text{NO GAPS}!}$:	If a moral theory is adequate, then there is no broad class of decision situations for which MOAC offers no meaningful guidance; in other words, it must not result in (a myriad of) systematic deontic gaps.

$C_{\neg \text{ADEQ}}$:	MOAC is not adequate.

The Challenge, then, ultimately relates to the Real Challenge as a symptom that has its underlying cause in a particular tactical move—the Consequentialist Standard Move (CSM)—that MOAC has made to manage

the Real Challenge: It would be unbearable if moac were unable to make genuine unconditional assessments in a widespread class of collective decision situations. However, in all interesting collective decision situations, they seem capable of providing only *conditional* assessments. To avoid this, the champions of moac are willing to allow the consideration of facts about the actual actions of the respective other agents to resolve the conditional assessments. The champions of moac, thus, *tacitly assume* that the original descriptions of the cases were incomplete, i.e., they implicitly reject $P_{\exists \text{NON-DEC}}$. Therefore, they pretend there were normatively relevant facts missing—namely, facts about how the agents actually act. This move, the addition of factual actions, however, allows us to find situations—namely, Troublemakers—in which the interdependencies are so unfavorable that one can find combinations of actions in which the actions have to be assessed as right, even though the involved agents collectively bring about suboptimal and even unbearable overall consequences. This is the Challenge: a consequence of the Real Challenge together with a quick and dirty fix, i.e., the csm.

At this juncture, it is worthwhile to revisit Regan's impossibility proof and consider it from this perspective (cf. Subsection 3.5.2). Let us return to Regan's step-by-step application of the csm, starting with a recall of his illustrative toy example (Regan 1980, p. 18), which is structurally analogous to Two Factories:

Case 3.7 (Whiff and Poof) *Suppose that there are only two agents in the moral universe, called Whiff and Poof. Each has a button in front of him which he can push or not. If both Whiff and Poof push their buttons, the consequences will be such that the overall state of the world has a value of ten units. If neither Whiff nor Poof pushes his button, the consequences will be such that the overall state of the world has a value of 6 units. Finally, if one and only one of the pair pushes his button (and it does not matter who pushes and who does not), the consequences will be such that the overall state of the world has a value of 0 (zero) units. Neither agent, we assume, is in a position to influence the other's choice.*

This case can be represented in the following normal form:

		Poof	
		not-push	push
Whiff	not-push	6	0
	push	0	10

Note that this description contains no assumptions regarding what Whiff and Poof actually do or will do. He starts from the case as given here.

Next, we return to what Regan called a "precise necessary condition for exclusive act-orientation" (Regan 1980, p. 114) that he calls "the partial definition". Recall that the property of exclusive act-orientation is the core piece of Regan's result:

> Any exclusively act-oriented theory must, in this example, on any assumption about Poof's (Whiff's) behavior, identify some non-empty subset of the set of acts comprising "pushing" and "not-pushing" such that Whiff (Poof) satisfies the theory if and only if he does some act from that subset.

Notice that, by the partial definition, every exclusively act-oriented theory is meant to deliver a *non-empty* subset of the option space of an agent. In a sense, this is a built-in commitment to $P_{\exists\text{GAPS}}$ for such theories. Regan gives an argument for this commitment (ibid., p. 115):

> If an exclusively act-oriented theory selected the empty subset on any assumption about Poof's (Whiff's) behavior, then it would direct Whiff (Poof) to do the impossible. (Selecting the empty subset is not the same as directing the agent not to push. "Not-pushing" is an act for our purposes. Selecting the empty subset is directing the agent to neither push nor not-push, which he cannot do.)

In other words, Regan argues that permitting empty sets of right actions would entail a violation of some version of the venerable principle that "ought implies can." If the relevant set were empty for a given decision situation and agent, MOAC would still require the agent to perform one of the actions in that set. Consequently, the agent would be obligated to perform *no action at all*—a logical impossibility, as such an action cannot be performed. *This* is precisely why Regan incorporated Resolvability into his partial definition.

Next recall Regan unfolding his central argument (ibid., p. 115):

> Suppose there is an adaptable theory T which satisfies the partial definition. Suppose further that Poof does not push. Since T satisfies the partial definition, there is some non-empty subset of the set of acts "pushing" and "not-pushing" such that Whiff satisfies T (while Poof does not push) if and only if he does an act from that subset. Call the subset S. We can deduce what S must be from the assumptions we have made about T. We know that Whiff satisfies T if and only if he does an act from S. So, if Whiff does an act from S, he satisfies T. Since T is adaptable, T [embraces MOCOR]. That means that any agent who satisfies T produces the best possible consequences in his circumstances. If Whiff produces best possible consequences in his circumstances, which include Poof's not-pushing, he must not-push. Therefore, if Whiff satisfies T, he not-pushes. Remembering what we have already established, that if Whiff does an act from S, he satisfies T, we can conclude that if Whiff does an act from S, he not-pushes. But remember

> also that *S* is non-empty. The only non-empty set such that if Whiff does an act from that set he not-pushes is of course the set consisting of the act "not-pushing". Therefore *S* consists of the act "not-pushing". In sum, if Poof does not push, then Whiff satisfies *T* if and only if he (Whiff) not-pushes also.

According to my analysis, things go wrong instantly, right at the second sentence, where Regan assumes that Poof does not push, adding this fact to Whiffs "circumstances". Actually, given Regan's (MOCOR-like) explications of the act-consequentialist criterion of rightness, Regan had to smuggle this assumption into his proof. Consider a version without it:

> Suppose there is an adaptable theory *T* which satisfies the partial definition. Since *T* satisfies the partial definition, there is some non-empty subset of the set of acts "pushing" and "not-pushing" such that Whiff satisfies *T* (while Poof does **???**) if and only if he does an act from that subset. Call the subset *S*. We can deduce what *S* must be from the assumptions we have made about *T*. We know that Whiff satisfies *T* if and only if he does an act from *S*. So, if Whiff does an act from *S*, he satisfies *T*. Since *T* is adaptable, *T* [embraces MOCOR]. That means that any agent who satisfies *T* produces the best possible consequences in his circumstances. If Whiff produces best possible consequences in his circumstances, he must **???**. Therefore, if Whiff satisfies *T*, he **???**. Remembering what we have already established, that if Whiff does an act from *S*, he satisfies *T*, we can conclude that if Whiff does an act from *S*, he **???**. But remember also that *S* is non-empty. The only non-empty set such that if Whiff does an act from that set he **???** is of course the set consisting of the act **???**. Therefore *S* consists of the act **???**. In sum, if Poof does **???**, then Whiff satisfies *T* if and only if he (Whiff) **???** also.

The problem here is that Regan's partial definition ensures the existence of a non-empty set *S* of right actions according to MOAC. However, no such set *S* exists given only Whiff and Poof. The set of right actions only 'becomes' non-empty once we add a fact about what the other agent does. This, however, creates a significant problem: at least one agent cannot satisfy *T* in any meaningful way. Specifically, whoever acts first inevitably performs an action that was not part of 'his' set *S*, because, for that agent, the set of right actions is empty. It is worth noting that Regan was fully aware of the need to introduce such an additional fact. Early in his book, he explicitly acknowledges this requirement (ibid., p. 18):

> Now, if we ask what AU [i.e., MOAC] directs Whiff to do, we find that we cannot say. If Poof pushes, then AU directs Whiff to push. If Poof does not push, then AU directs Whiff not to push. Until we specify how Poof behaves, AU gives Whiff no clear direction. The same is true, *mutatis mutandis*, of Poof.

However, rather than using this observation to build an argument against MOAC directly, Regan focused on demonstrating that MOAC fails with respect to PMH.

Given all this, I believe Regan's overall argument is best understood as follows: For a moral theory to truly qualify as exclusively action-oriented, the moral status of actions must depend solely on the nature of those actions themselves. Additionally, a moral theory should be generally applicable, meaning it must be capable of guiding agents' choices under ordinary circumstances, at least by default. As mentioned above, it seems deeply unsatisfactory to tell Ann and Ben in Two Factories that MOAC has nothing to say about their choices simply because they are in a collective decision situation and do not face 'true' individual decision situations. Furthermore, a moral theory must never demand the impossible of agents; specifically, there must always be an action that, if performed, satisfies the theory. Troublemakers, however, reveal a theoretical *trilemma* for proponents of MOAC (and indeed any exclusively action-oriented moral theory):

- They must either relinquish general applicability, i.e., abandon Deontic Completeness, by admitting that the theory has nothing to say in such cases;
- Or they demand the impossible, requiring agents to perform a right action without being able to identify one, thereby violating No Moral Dilemmas;
- Or they add facts about what other agents do or will do to the context of the decision, thereby violating Methodological Indeterminism.

Crucially, in the latter case, the theory fails to satisfy PMH, which remains an unacceptable option for proponents of MOAC. This conundrum—the Challenge in its internal form, Internal Challenge—leaves MOAC irredeemably untenable.

The main point I make in this chapter is then this: the Challenge as Internal Challenge, given closer inspection, does not *really* work as planned because there are no genuine violations of the PMH. As the agents, according to this quickly painted picture, act first and are assessed in retrospect, they can neither follow morality nor satisfy moral theories when acting, nor can they fail to do so. The challenge, thus, lies deeper: the core idea of PMH just runs empty—due to the deontic Gaps that were never filled meaningfully and substantially but were only hastily patched over argumentatively.

In light of all this, it becomes obvious what camp MOAC ultimately needs: A way to resolve the Real Challenge *without* running into any version of the Challenge, i.e., in a way that allows MOAC to fill the deontic gaps,

to give unconditional assessments that are, actually, in line with basic (consequentialist) intuitions. This is what remains to be done and what I strive to do in the further course of this thesis.

Chapter 7

Of New Consequences

This chapter aims to illustrate and spell out my central approach, which I will call the "Intermediate Outcomes Approach" (or just: the "Approach"). I discuss how it can help to actually fill the identified deontic gaps by exploiting newly discovered consequences of arbitrary actions in collective decision situations. The result is a *multi-agent* version of consequentialism equipped with the principled capability to provide non-conditional assessments for interesting collective decision situations, including Troublemakers. Along the way, we will find a new, *unified representation* of collective decision situations that fits well with the Approach and allows us to visualize both the Challenge and the Real Challenge. This will make us realize that there are actually several ways of exploiting the newly discovered consequences. I will call these *multi-agent amendments*, and I will present a selection of possible amendments that I take to contain the most interesting and important candidates. At the end of this chapter, there remains only one decision left: which amendment moac should embrace—and why. This, then, will be the topic of the next and last substantial chapter.

7.1 New Grounds for Consequentialism

At first glance, the matter might look hopeless. To solve the Real Challenge, moac must be able to give non-conditional assessments of all the agents' options (or at least of their 'right' options) in interesting collective decision situations. But in order to stay true to its act-consequentialist roots, it must not resort to anything other than the consequences of these options. Yet interesting collective decision situations—and thus Troublemakers—are characterized precisely by the fact that the actual consequences of actions depend essentially on what the other agents do. Hence, Compositionalism was also, to all appearances, a hopeless position, while the Genuine Kind View prevailed. We recall:

Claim 5.1 (Compositionalism) All *collective decision situations can be reduced to individual decision situations (plus some structure).*

Claim 5.2 (Genuine Kind View) *Some collective decision situations cannot be reduced to individual decision situations (plus some structure).*

How can the consequentialist extricate himself from this apparently hopeless situation?

The answer is: She must defend Compositionalism and thus deny the existence of non-decomposable and, hence, of allegedly interesting collective decision situations. For this, she 'merely' has to show how apparently interesting collective decision situations could be decomposed into individual decision situations. What MOAC lacks for this are, obviously, consequences that can be directly assigned to individual actions without conditionalizing on other agents' actions. But from where to take such consequences?

The answer is *so* evident that one can only say that MOAC did not see the forest for the trees. For, of course, we have long since seen the relevant consequences and have even listed them quite explicitly several times in the preceding chapters. In the following, I will mark the forest and then defend that this is not a 'lunatic insight'.

7.1.1 "Like Scales Fell From His Eyes…"

At this point, we are in search of overlooked consequences. This first raises the question of what we are actually looking for when we search for consequences. The following rough characterization will probably suffice to give us a reasonably reliable criterion for what counts as a consequence of some given action ϕ: We are looking for something that would be the case if ϕ were performed and would be absent if, instead of ϕ, some other action were performed. This should remind us of Jackson's

Principle 3.1 (Difference Principle) *The morality of an action depends on the difference it makes; [i.e.,] it depends on the relationship between what would be the case were the act performed and what would be the case were the act not performed.*

So, let's return to our running example to look for such differences. Recall:

Case 3.1 (Two Factories) *Ann and Ben each own a factory along the same river. Both can choose to produce either cleanly or cheaply, thereby polluting the river. The local market is highly competitive. Thus, a factory that produces cleanly would become noncompetitive if the other factory pollutes. The local social system is underdeveloped, and the economic situation is terrible. Hence, if a factory closes, this will cause significant unemployment and social hardship.*

If at least one factory produces cheaply, the resulting pollution will eventually destroy the local ecosystem and erode the livelihood of a village downstream. However, any additional polluter would not make the situation worse in this regard. Ann and Ben decide and act independently in the sense that neither of them can coordinate with the other, nor can any agent observe the actions of the other. Thus, whatever each agent does, they would do it regardless of what the other does.

The straightforward representation of the Troublemaker in terms of a normal form was given by this normal form:

		Ben	
		pollute	produce cleanly
Ann	pollute	second-worst	worst
	produce cleanly	worst	best

The essential question now is this: What difference does it make if Ann produces cleanly? The answer is: it changes the decision situation that remains for Ben! That insight, in a way, has been the whole idea behind conditionalization. Suppose

Fact 7.1 *Ann produces cleanly*

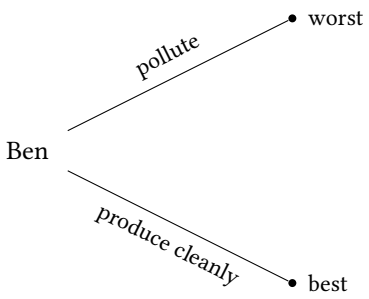
then we have ascertained that

(44) It is right for Ben to produce cleanly

and

(45) It is wrong for Ben to pollute

are true relative to $[\![\text{Two Factories} \oplus \text{Fact 7.1}]\!]$. We have already established that Fact 7.1 allows us to reduce Two Factories to this individual decision situation of Ben:



However, what would happen if Ann acted differently? Let us entertain

Fact 7.2 *Ann pollutes.*

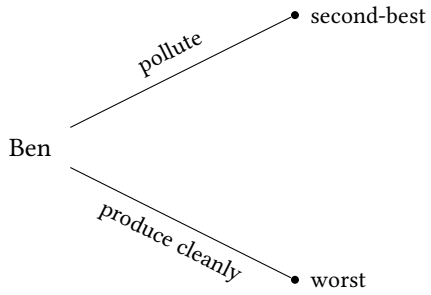
Relative to [[Two Factories ⊕ Fact 7.2]] we get

(46) It is right for Ben to pollute

and

(47) It is wrong for Ben to produce cleanly.

Furthermore, Fact 7.2 allows us to reduce Two Factories to this individual decision situation for Ben:



Thus, if we are looking for differences Ann's actions make, then these two remaining individual decision situations of Ben, the results of the reduction of Two Factories relative to Ann's respective action, are pretty plausible candidates. The other way around, of course, applies to Ben as well: if he acts first, this brings about one of two possible remaining individual decision situations for Ann. Looking at the normal form of the case, we can express this insight like this: if Ann acts first, the consequence of her action corresponds to the row associated with her action; if Ben acts first, the consequence of his action corresponds to one of the two columns.

This insight might very well help us solve both the Real Challenge *and* the Challenge. If we accept these outcomes as proper outcomes *and* if we find a way to utilize them appropriately, then we might fill the gaps in a way that does not reiterate the Challenge. In the remainder of this thesis, I turn to why we should accept these outcomes and how to utilize them appropriately. But for now, let's concentrate on the general idea first.

Two questions may arise: first, how should we assess the later-acting agents' actions, and second, what should we say in the rather esoteric edge case of simultaneous action? The answer to the first question is particularly straightforward in the two-agent case. Whoever acts second is in the remaining individual decision situation brought about by the first acting agent. To

assess this situation is not a challenge, but the daily business of MOAC and MOCOR. In cases with more agents, we can just reiterate the procedure: The result of the action of the second-acting agent is a further reduced decision situation. We can recursively apply the same procedure until, finally, we end in a classical individual decision situation of the last-acting agent.

Regarding the synchronous case, things also seem pretty straightforward: Since the actual context of their decision cannot contain any fact of what the other agent does up to the moment of acting (for methodological reasons, recall Methodological Indeterminism), both bring about the individual decision situation of the other agent. Even though the other agent's decision situation is immediately resolved in this case of synchronous action, this does not affect the assessment in the original context.

Call this approach, i.e., considering the remaining decision situation of the other agent (or agents) as the consequence of an agent's actions within a collective decision situation, the Intermediate Outcomes Approach ("the APPROACH", in short). It means breaking down the one-step approach underlying our current understanding of collective decision situations into a step-by-step approach: Instead of just assigning *final outcomes* to combinations of actions, we start with a first action that reduces the collective decision situation to some intermediate outcome that is itself a decision situation. If more than two agents are involved, a sequence of further action-by-action reductions finally leads to an individual decision situation before the last action leads to a final outcome.[132]

It instantly follows that we can maintain Compositionalism. Recall

Claim 5.1 (Compositionalism) All *collective decision situations can be reduced to individual decision situations (plus some structure).*

Here is how we can decompose arbitrary collective decision situations into individual ones. Let D be a collective decision situation represented by $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$. Then, we can decompose D in $n := |\mathcal{A}|$ individual decision situations D_{A_1}, ..., D_{A_n}, one for each agent, with the structure

$$\mathsf{D}_{A_i} := \left\langle A_i,\ \Phi_{A_i},\ \mathsf{Out}_C : \Phi_{A_i} \to \mathcal{O}_D^{A_i} \right\rangle,$$

where

$$\mathcal{O}_D^{A_i} := \{\, D_{\downarrow\phi} \mid \phi \in \Phi_{A_i} \,\}.$$

Note how all these decision situations are structurally interconnected because the outcomes of each of them are reduced versions of the original collective decision situation D, involving all the other agents and their remaining choices. Let's call the set $\mathfrak{I}_D := \{\, \mathsf{D}_A \mid A \in \mathcal{A}_D \,\}$ D's *decomposition*.

[132] Note that we have restricted this investigation to *finite* collective decision situations.

So it turns out that the so attractive and formerly apparently hopeless claim of Compositionalism prevails over the Genuine Kind View after all. Recall that Compositionalism is so attractive for MOAC because, if true, it would unlock the universal applicability of MOCOR in the collective domain—given that the newly discovered consequences can be ranked as required by MOCOR. This should make the champions of MOAC optimistic.

This brings us back to the central task that remains to be carried out. For APPROACH to be ultimately useful for MOAC, it must be decided how the newly discovered consequences, i.e., the reduced decision situations, are to be integrated into MOAC's overall framework. This comes down to deciding how these decision situations are to be evaluated (or at least ranked) morally and, thus, to an axiological question in the broad sense.[133] However, this question of the appropriate moral assessment of decision situations as consequences is anything but trivial, as different amendments have to be considered and, somehow, compared.

In the remainder of this chapter, I aim to demonstrate that the approach I propose is far from far-fetched and is, in fact, readily integratable into broader consequentialist thought. Following this, I will revisit PMH to examine how it might (or might not) guide us in addressing the moral assessment of decision situations. Lastly, I will present several potential candidates for amendments. The task of evaluating and comparing these amendments—as well as establishing the criteria for such comparisons—is reserved for the next chapter.

7.1.2 Exotic and Esoteric? Or Old Wine in New Bottles?

MOAC is a consequentialist *moral* theory. However, there are other, non-moral kinds of consequentialist theories for which the idea of decision situations being consequences is prevalent (for an overview of the different types of normative theories at play here, see Figure 7.1). For instance, classical decision theory, understood as a theory of instrumental rationality, has long recognized a nuanced picture of kinds of outcomes. Most importantly, the von Neumann–Morgenstern utility theorem, one of the fundamental theorems of utility theory, lets agents choose between *lotteries*, scenarios with

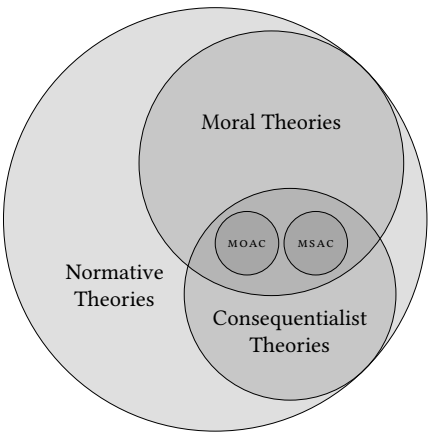


Figure 7.1: How normative, moral, and consequentialist theories relate to each other. Standard decision theory, as a theory of instrumental rationality, is a non-moral consequentialist theory.

[133] Recall Subsection 2.3.3 on page 33.

uncertain outcomes, some of which might be lotteries themselves.[134] The utility of these lotteries is then thought of as a function of the possible *final* outcomes and the probabilities with which they obtain, with the function typically being their expected utility. In this sense, lotteries can be considered a kind of *intermediate* outcomes.

Game theory, with its intrinsic focus on the multi-agent dynamics of decisions and actions, develops the idea of decision situations as (intermediate) outcomes even further. As agents maneuver within games, the choices of one agent are typically equivalent to selections of sub-games for other agents. From the perspective of game theory, an extensive form, as seen many times in this thesis already, is a tree (or at least a directed acyclic graph) representing the (remaining) game, where transitions correspond to the agents' actions. Every action corresponds to the selection of a sub-tree, and the consequences of the actions correspond to the remaining game and, thus, to the decision the following agent has to make.

The readiness to accept decision situations as outcomes is not restricted to the consequentialist domain of instrumental rationality, though. Many subjective consequentialist theories like MSAC are rooted in expected utility theory and, more generally, decision theory. Recall

Principle 3.9 (MSCOR (prototypical)) *It is right to perform a certain action if and only if there is no alternative with expectedly better consequences.*

Such theories can easily consider scenarios where the actions' immediate consequences are lotteries. Recall the original version taken from Jackson (1991, p. 462):

Case 3.8 (The Drug) *Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the literature has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it. One of the other two drugs, either B or C, will completely cure the skin condition; the other though will kill John, and there is no way that she can tell which of the two is the perfect cure and which is the killer drug.*

In Chapter 3, we represented The Drug as a decision tree (cf. Figure 7.2), featuring three lotteries. However, the case can be easily reformulated into a *sequence* of two subsequent decisions, resulting in a scenario that is essentially equivalent to the original, without posing any new challenges for camp MSAC. Consider

[134] If the uncertainty can be expressed in terms of probabilities such decisions are typically described as decisions under *risk*. This is distinct from decisions under *uncertainty*, where outcomes lack specific probabilities.

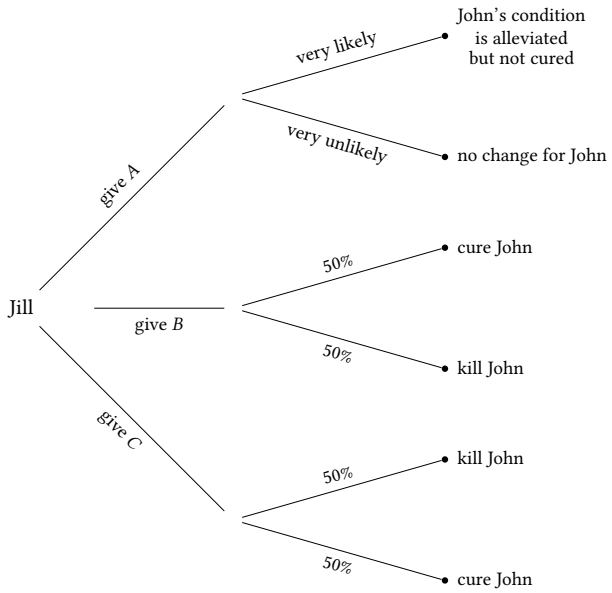


Figure 7.2: The extensive forms of The Drug.

Case 7.1 (The Drug (II)) *Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the literature has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it, drug B will completely cure the skin condition, and drug C will kill John. Drug A is on the tray directly in front of Jill, while drugs B and C are in the cabinet in the next room. However, the labels of these two drugs in the cabinet are illegible due to the passage of time, and there is no way for Jill to figure out which one is which. As a matter of fact, but unbeknown to Jill, the drug standing on the left is drug B, and the drug on the right is drug C.*

If Jill decides against drug *A*, she decides to go to the cabinet, where she faces another decision. We may assume that if Jill had to pick between the two, she would select randomly with a probability of 50%. It is easy to see that Jill's new decision is very similar and open to the same analysis and assessment as the original The Drug case (cf. Figure 7.3). But in The Drug (II), Jill can either decide on an action with a lottery outcome (giving the suboptimal drug *A*) or bring herself into another decision situation. This should not bother the subjective consequentialist at all. And lotteries and decision situations are not so different after all, as decision theory demonstrates—and Kagan's (2011) discussion of the Challenge demonstrates, for instance, that subjective consequentialists see it this way too (cf. Subsection 4.4.1).

For object consequentialism, however, it seems to remain at least an unusual conceptualization to think of agents' actions as lotteries—or, more

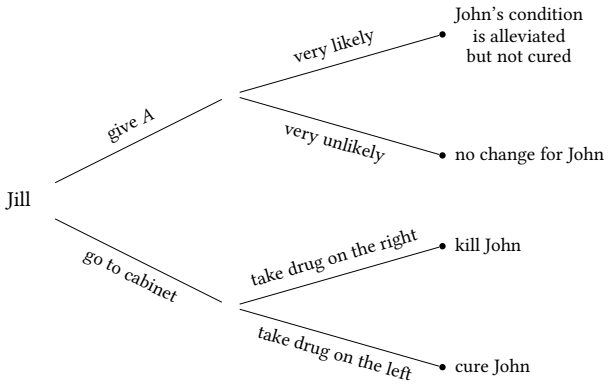


Figure 7.3: The extensive forms of The Drug (II).

specifically, selections of other agents' decision situations. Historically, objective consequentialism has focused primarily on the singular choices of individual agents *without uncertainty*. There was not much space for exogenous variables. And if they were considered, then typically not so much as being dynamic elements of change, but rather as static and pre-determined. Though providing clarity in specific scenarios, such an approach inadvertently narrows the consequentialist lens, omitting the more intricate tapestry.

This is certainly not to suggest that *no one* in camp MOAC has considered decisions involving genuine uncertainty. Such an omission would be highly surprising, especially given the challenge posed by Prior (cf. Prior and Raphael 1956; see above). Indeed, numerous authors have engaged with the (possible) indeterminacy or under-determinacy of future 'developments' in the context of objective consequentialism (cf. Horty 2001; Sinnott-Armstrong 2022). The actualism–possibilism debate itself assumes the possibility of future decisions explicitly. Prior, for his part, proposed using objective probabilities to evaluate uncertain futures, an approach that continues to find both defenders[135] and critics (cf. Wroński 2020). I will examine this approach in greater detail later.

My point is this: The historical trajectory is not merely a theoretical curiosity but represents a persistent methodological gap. The importance of the underdetermination of the future has yet to permeate the mainstream of objective consequentialism. However, when objective act-consequentialism is applied to the complex dynamics of multi-agent interactions, its traditional focus on individual decision situations with clearly defined consequences for each option appears overly simplistic. In collective decision situations, it becomes evident that the context in which one agent makes a decision can be significantly shaped—or even determined—by the decisions of others.

[135]For example, Jackson explored a probabilistic solution to the actualism–possibilism challenge, deliberately leaving the specifics of the relevant account of probabilities open (cf. Jackson 2014).

Objective consequentialism must create conceptual space for uncertainty. To paraphrase John Donne (1923): *No agent is an island.* This interconnectedness implies that, while agents must be regarded as free in their choices (in line with Methodological Indeterminism), their mutual dependence on what can be collectively achieved gives rise to a dynamic that defies static frameworks. In this way, the reality of collective interdependence underscores the need for a rigorous objective consequentialist framework to embrace uncertainty as a fundamental aspect of multi-agent scenarios—i.e., collective decision situations.

So, the idea underlying the Approach, accounting for this kind of non-epistemic but purely theoretical-methodological uncertainty by incorporating decision situations as intermediate outcomes in moac's framework, is by no means an esoteric or exotic approach. Instead, it corresponds to the already-lived consequentialist practice, well-anchored in preliminary work on decision theory. In order to get a better understanding and feeling for the implications of Approach, it is worth to next develop a new way of representing collective decision situations under the perspective of the Approach.

7.2 Towards a Unified Representation: The Generalized Extensive Form

To systematically examine how moac might exploit the freedom revealed by Approach, a more informative representation of Coordination Cases is helpful. This representation should explicitly carve out a space for assessing the novel intermediate outcomes proposed by Approach, aiming to offer a cohesive visualization of decision situations. Furthermore, it would be welcome if Sequential Cases could also be represented immediately in a corresponding formal structure. Then we would have a structure that might allow us to find a solution to both the Real Challenge and the Challenge for Coordination Cases as well as for Sequential Cases in one go.

While the normal form representation simplifies matters for elementary cases, its general applicability is limited. Given the Approach, initial actions are akin to selecting rows or columns in a two-agent scenario, yielding intermediate outcomes. Subsequent actions then pinpoint final outcomes. This 'zooming-in' perspective on outcomes proves insightful for simple cases, but attempting to visualize scenarios with more than three agents quickly becomes convoluted. For instance, visualizing a three-agent scenario is analogous to the gradual dissection of a cube, and the complexity grows with additional agents, each introducing an additional dimension. Although the matrix-like normal form offers computational advantages—similar to adja-

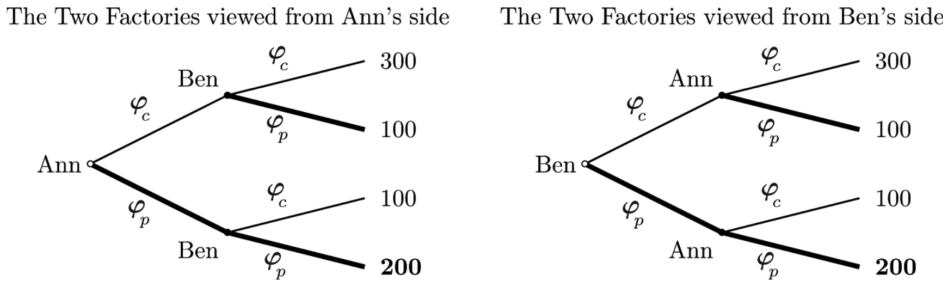


Fig. 2.—The Two Factories as extensive form game

Figure 7.4: Pinkert's non-unified representation of his Two Factories case, illustrated with two *possible* extensive forms that he seemingly treats as a single "extensive form game" Pinkert 2015, p. 975. Note that additional trees would be needed if synchronous actions were permitted and all possibilities were accounted for.

cency matrices as representations of graphs—it quickly loses intuitive appeal and becomes hard to grasp.

Historically, the challenge with using the extensive form for Coordination Cases has been its reliance on depicting outcomes within *a single* graph. The indeterminacy in the order of actions introduces inherent ambiguity to such representations. To retain generality, one might consider constructing decision trees for *all* possible action sequences, resulting in a 'forest' of extensive forms (see Figure 7.4). However, with Approach providing all necessary intermediate results, it is now possible to disentangle and consolidate these potential sequences into a single, unified graph. This section aims to introduce a *generalized extensive form* designed to navigate the complexities of the MOAC framework with clarity and precision.

In achieving this, retaining simplicity and clarity in the extensive form without omitting MOAC-relevant information is vital. For comparison, consider how the normal form streamlines Coordination Cases. Each cell in this format embodies multiple potential outcomes, represented singularly if the outcomes share values and result from combinations of the same actions. For a clearer perspective, consider the Two Factories example: Compare the outcomes of sequences where Ann pollutes before Ben and vice versa. Though technically distinct (there are facts that are true relative to the outcome of the first sequence and false relative to the second, for instance, that Ann acted first), their representation is condensed into a single cell due to their identical nature in the order-invariant Two Factories scenario. This suggests the following[136] definition:

[136]This definition uses the notion of a proper part of a combination, i.e., the relation $\sqsubset$ as defined in Subsection 5.4.1 on page 169. Further, it makes use of the formerly introduced notions related to the equivalence relation $\sim^*$ to perform a primitive version of *state lumping* (Kemeny and Snell 1960), cf. Subsubsection 5.2.1.2, page 155.

Definition 7.1 (Generalized Extensive Form)
Let D be a maximal and order-invariant Coordination Case with representation $\mathsf{D} = \langle \mathcal{A},\ \Gamma, \mathsf{Out}_C :\ \Psi\ \to \mathcal{O} \rangle$ (where C is the actual context of D) and $|\mathcal{A}| = n$.

The generalized extensive form $\mathcal{G}(D)$ *of D is a directed acyclic graph, i.e., a tuple $\mathcal{G}(D) = \langle \mathcal{S},\ \mathcal{E},\ S_\emptyset \rangle$ with a set of states $\mathcal{S}$, a set of transitions $\mathcal{E}$, and initial state $S_\emptyset \in \mathcal{S}$. These sets are defined as follows:*

(i) The set of states $\mathcal{S}$ *includes a state for every outcome, be it final or intermediate, from a consequentialist perspective. Given an arbitrary set of representatives $\mathcal{R}$, we first define the set of* final outcome states *as*

$$\mathcal{S}_n\ :=\{\ S_\Upsilon \mid \Upsilon \in \mathcal{R}\ \}.$$

Next, we define recursively

$$\mathcal{S}_{i-1} = \bigcup_{S_\Upsilon \in \mathcal{S}_i} \left\{\ S_{\Upsilon \ominus \phi} \mid \phi \in \Upsilon\ \right\}$$

down to $\mathcal{S}_0 = \{S_\emptyset\}$, the initial state singleton. *Finally, we define the set of states as the union of these sets, i.e.:*

$$\mathcal{S}\ :=\bigcup_{i=0}^{n} \mathcal{S}_i.$$

(ii) The set of transitions $\mathcal{E}$ *is a set of (directed) edges that, all together, represent all sequences of actions by which D could be resolved. For $0 \leq i < j \leq n$, we define*

$$\mathcal{E}_{i,j}\ :=\left\{\ \langle S_\Upsilon, S_{\Upsilon'} \rangle \mid S_\Upsilon \in \mathcal{S}_i,\ S_{\Upsilon'} \in \mathcal{S}_j,\ \Upsilon \sqsubset \Upsilon' \right\}.$$

Each of these sets of edges represents all the transitions from all states $S_\Upsilon \in \mathcal{S}_i$ to the corresponding states $S_{\Upsilon'} \in \mathcal{S}_j$. Thus, by construction, each edge represents the combination of action Υ'' with $\Upsilon' = \Upsilon \oplus \Upsilon''$ by the corresponding agents (where $|\Upsilon''|$ might be 1). Finally, we define the set of transitions as the union of all of these sets, i.e.,

$$\mathcal{E}\ :=\bigcup_{i=0}^{n-1}\bigcup_{j=i}^{n} \mathcal{E}_{i,j}.$$

General extensive forms (shorter: GEFs) can be interpreted as follows. Let D be a collective decision situation and let $\mathcal{G}_D = \langle \mathcal{S},\ \mathcal{E},\ S_\emptyset \rangle$ be its general extensive form. We can make the following observations.

$S_\emptyset$ represents the initial state of D, and every path from $S_\emptyset$ to some $S_\Upsilon \in \mathcal{S}_n$ represents at least one proper combination. In fact, each such path represents

all proper combinations $\Upsilon' \in [\Upsilon] \subseteq \Psi$. All other states in between, i.e., all states $S_{\Upsilon''} \in \mathcal{S}_i$ (for $0 < i < n$) represent starting points of reduced decision situations after Υ'' (which is, by construction, guaranteed to be a proper part of a proper combination) has been performed. (Note that for each such proper part, there is exactly one corresponding partial path, even though multiple such paths might converge and merge again, as described above.)

Accordingly, every edge in $\mathcal{E}$ corresponds to one or more actions that can happen during D's unfolding. While an edge between two states $S_\Upsilon \in \mathcal{S}_i$ and $S_{\Upsilon'} \in \mathcal{S}_{i+1}$ represents a single action, an edge between two states $S_\Upsilon \in \mathcal{S}_i$ and $S_{\Upsilon'} \in \mathcal{S}_j$ with $j - i > 1$ represents the synchronous performance of $j - i$ actions.

Finally, every state $S_\Upsilon \in \mathcal{S}$ can be seen as the 'defining anchor' of a sub–decision situation within D with a corresponding general extensive form that is a subgraph of the original one. Let $\Upsilon \in \Psi_D$ be a proper combination. Then, for every proper part $\Upsilon' \sqsubset \Upsilon$ of that combination, D can be reduced to $D_{\downarrow\Upsilon'}$. Then we can infer $\mathcal{G}_{\Upsilon'} = \langle \mathcal{S}_{\Upsilon'},\ \mathcal{E}_{\Upsilon'},\ S_{\Upsilon'} \rangle$ as

$$\mathcal{S}_{\Upsilon'} = \{\ S_{\Upsilon''} \mid \Upsilon' \sqsubseteq \Upsilon''\ \}$$

and

$$\mathcal{E}_{\Upsilon'} = \{ \langle S_{\Upsilon''}, S_{\Upsilon'''} \rangle \in \mathcal{E} \mid S_{\Upsilon''}, S_{\Upsilon'''} \in \mathcal{S}_{\Upsilon'} \}.$$

Put simply, to construct $\mathcal{G}(D_{\downarrow\Upsilon'})$, we remove all parts of $\mathcal{G}(D)$ that are inconsistent with Υ' having already been performed. Consequently, there is a one-to-one correspondence between the intermediate states and the individual decision situations that function as intermediate outcomes according to Approach.

It is helpful to introduce a shorthand for the set of all intermediate outcomes. First, given a GEF $\mathcal{G} = \langle \mathcal{S},\ \mathcal{E},\ S_\emptyset \rangle$, we observe that the *set of intermediate states*, $\mathcal{S}^{\text{inter}}$, is defined by construction as:

$$\mathcal{S}^{\text{inter}} = \bigcup_{i=1}^{n-1} \mathcal{S}_i.$$

In other words, starting with a GEF, we discard the initial state $S_\emptyset$ and all final outcomes—i.e., those states without outgoing edges.

Next, we define the *set of intermediate outcomes* as:

$$\mathcal{O}^{\text{inter}} = \{\ D_{\downarrow\Upsilon} \mid S_\Upsilon \in \mathcal{S}^{\text{inter}}\ \}.$$

Even for small Coordination Cases, GEFs can become quickly quite large and cluttered. As an example, Figure 7.5 illustrates a GEF for a minimal case: a maximal and order-invariant Coordination Case involving two agents and two options each. Traditionally, we would represent such a Coordination Case in normal form as illustrated in Figure 7.6.

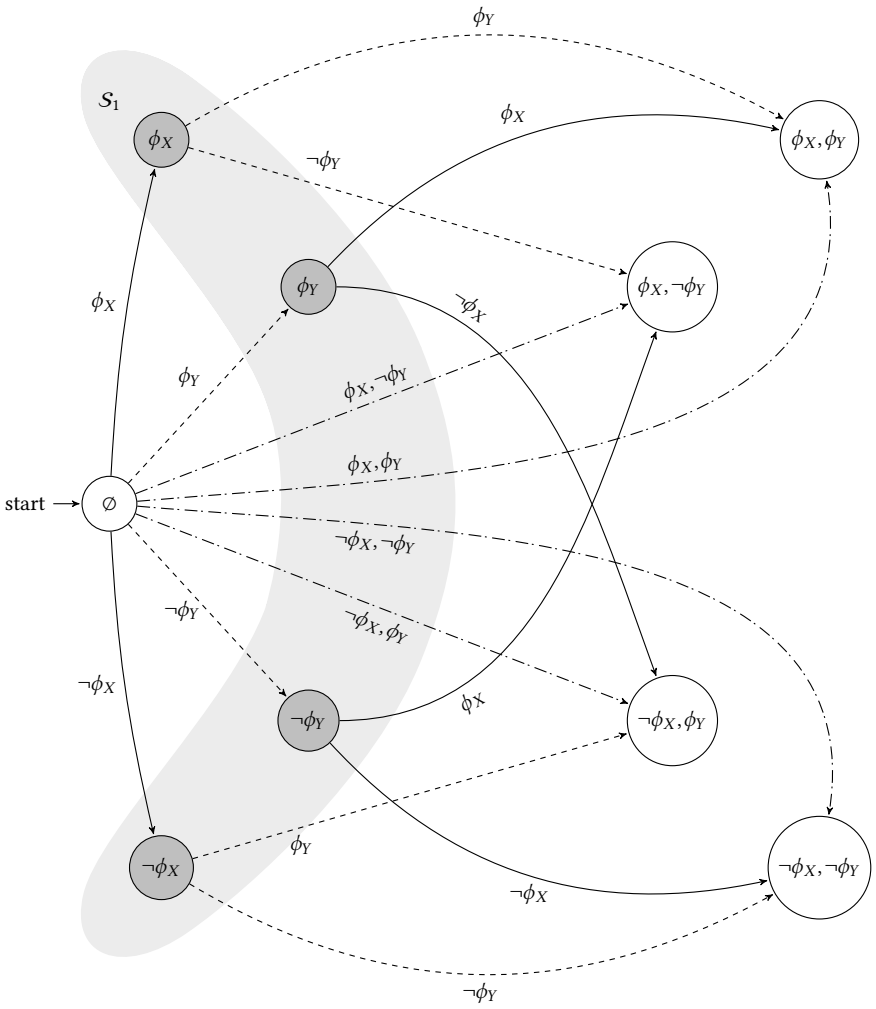


Figure 7.5: Generalized extensive form for a maximal, order-invariant Coordination Case with two agents (X and Y) each having two options (ϕ_z and $\neg\phi_z$ for $z \in \{X, Y\}$). Solid arrows denote actions by X; dashed arrows represent actions by Y; and dash-dotted arrows indicate synchronous actions by both. States are denoted by their index, e.g., state S_Υ is labeled "Υ". White-background states (excluding the initial state) show 'traditional' results such as $\mathsf{Out}(\langle \phi_X, \phi_Y \rangle)$, lumped following the above-defined procedure (note that, otherwise, we would have eight such white states). Dark gray states are intermediate, resulting from either X's row choice or Y's column choice, collectively forming the set S_1 (shown within the boomerang-shaped light-gray zone). These states are starting points for four sub-graphs (see Figure 7.7), resulting from actions corresponding to the edge leading to these states.

		Y	
		ϕ_Y	$\neg\phi_Y$
X	ϕ_X	$\mathsf{Out}(\phi_X, \phi_Y)$	$\mathsf{Out}(\phi_X, \neg\phi_Y)$
	$\neg\phi_X$	$\mathsf{Out}(\neg\phi_X, \phi_Y)$	$\mathsf{Out}(\neg\phi_X, \neg\phi_Y)$

Figure 7.6: Traditional normal form representation of a minimal Coordination Case with two agents and two options each, corresponding to the GEF shown in Figure 7.5. This compact representation highlights the strategic interactions but lacks the granularity of intermediate states and outcomes provided by the GEF.

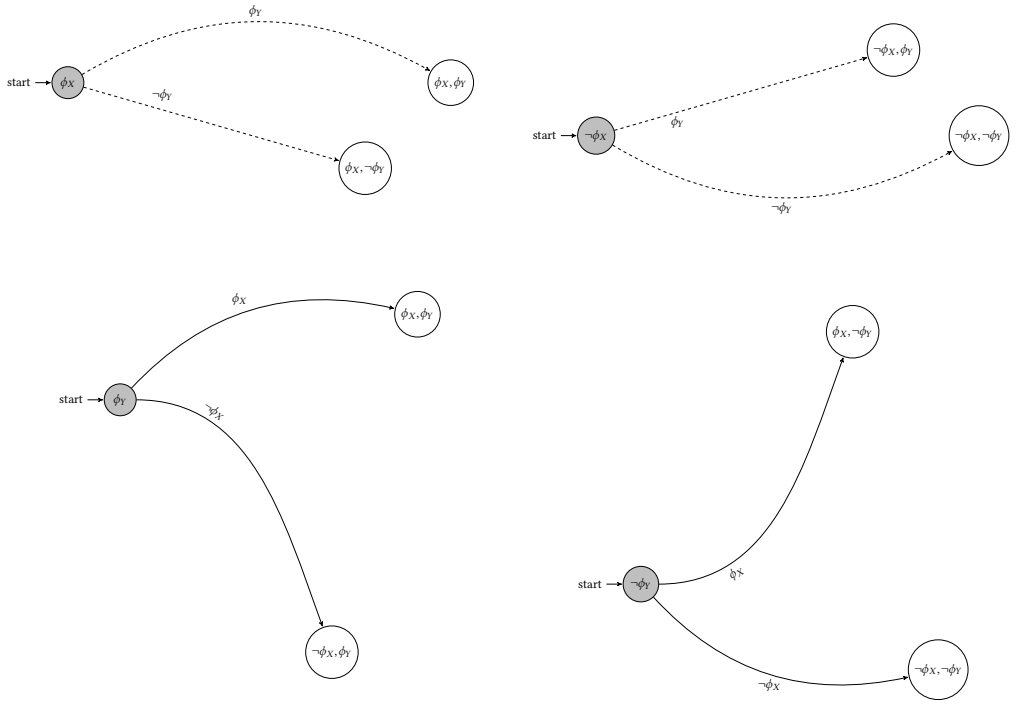


Figure 7.7: The four sub-graphs corresponding to $D_{\downarrow\phi_X}$, $D_{\downarrow\neg\phi_X}$, $D_{\downarrow\phi_Y}$, and $D_{\downarrow\neg\phi_Y}$, respectively (from left to right and top to bottom), and thus to one of the 'new outcomes' according to Approach.

One can see that the GEF of this case allows us to distinguish more clearly the relationships between the different combinations of actions and their occurrence (successive or simultaneous). Also, we can see at a glance the intermediate outcomes (cf. Figure 7.7), which are the same kind of thing as the GEF of the collective decision situation, namely a (sub–)graph.

At this point, thus, we can begin to *reason morally about paths* from the initial state to the final outcomes when thinking about where morality *should* 'guide' (in a sense still to be defined in more depth) the agents within a collective decision situation. Let us consider the following generic Two Factories–like case that we get from the above case by adding a fitting value profile:

		Y	
		ϕ_Y	$\neg\phi_Y$
X	ϕ_X	$-$	$---$
	$\neg\phi_X$	$---$	$+++$

We can add these valuations directly to the GEF (cf. Figure 7.8, left side, and for Job Market in Figure 7.9[137]). In the next step, we can then try to apply

[137]For Sequential Cases the concept of GEFs doesn't change much but we can easily see that we now have a truly unified representation for both kinds of cases.

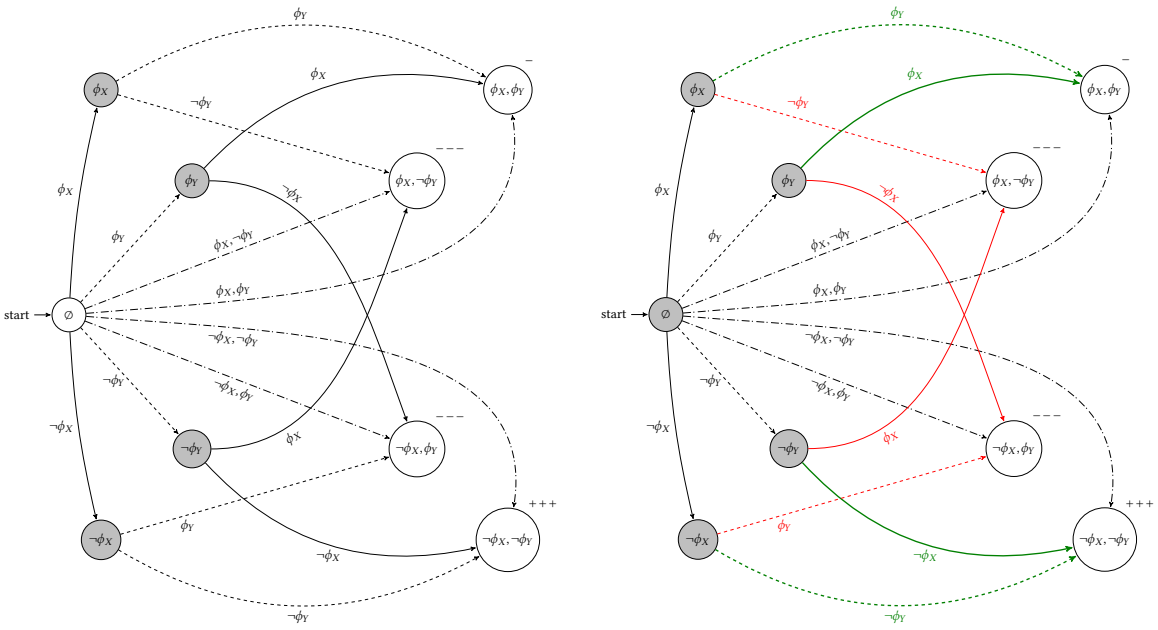


Figure 7.8: The GEF of Two Factories–like cases with value annotations (added at the top right of the states representing final outcomes, i.e., white states). On the right with additional moral assessment annotation (based on MOCOR): green for edges representing right actions and red for edges representing wrong actions. For actions corresponding to gray actions, MOCOR is not yielding assessments, i.e., only future-involving conditional assessments are available. (In addition, we can derive individual backward-looking assessments in Pinkert's style, cf. Subsubsection 3.5.2.3, page 98.)

MOCOR. This gives us a visualization of both the CHALLENGE and the REAL CHALLENGE (cf. Figure 7.8, right side). We recognize the CHALLENGE in the fact that there are green edges towards the ϕ_X,ϕ_Y state. As soon as we end up in that state and we start the counterfactual reasoning typical for the CSM, we find—correctly—that if any of the agents had acted differently, they would only have made things worse, i.e., acted wrongly and, accordingly, followed a red edge. The CHALLENGE, as analyzed in the last chapter, thus has its origin in that the champions of MOAC, to get assessments at all, move right to the final state and then, when arguing for the rightness of the two actions that lead there, only moves back to the ϕ_X or ϕ_Y state (depending on whether they are pondering on ϕ_Y or ϕ_X).

The REAL CHALLENGE is visualized to such a degree that we now recognize that a violation of PMH would happen only if we found a green path from the initial state to the ϕ_X,ϕ_Y state. But no such path exists, simply because *no green edge leaves the initial state at all.* The REAL CHALLENGE is precisely this: It is not the case that a green path leads to a suboptimal state; instead, there is no completely green path leading anywhere; in particular, there is no green path to the optimal state. According to MOAC, there simply

is *no right first step* in any direction. Thus, there is no possibility of following the commandments and recommendations of MOAC (as here there are none). All there is: are deontic gaps.

Before I move on and discuss how consequentialists might close these gaps, we can reconnect to the end of the first part: It is now easy to explain what was wrong with Jackson's reasoning in Subsection 4.4.3: the *frame of reference* just wasn't right. Let me explain.

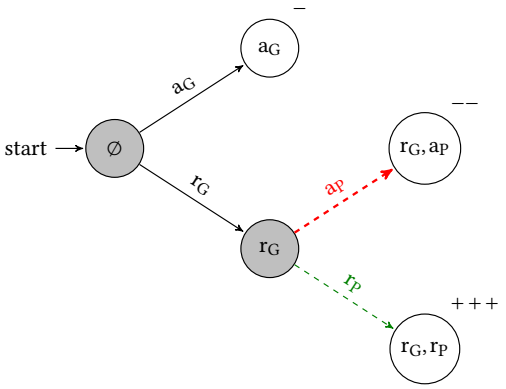


Figure 7.9: The (unspectacular) GEF of Job Market with annotations where "a" stands for "accepting the job offer", "r" stands for "rejecting the job offer", "G" stands for "George", and "P" stands for "Paul". Extensive forms just remain extensive forms. Yet, we see that also, in this case, we only get assessments for Paul's (hypothetical) decision situation—and that it seems more than plausible to say that George's option of rejecting the job has an individual decision situation as a consequence, namely Paul's. Arguably, the same line of reasoning that resolves Coordination Cases should resolve Sequential Cases automatically as Coordination Cases are, in light of the APPROACH, just *superpositions* of several Sequential Cases and solving a Coordination Cases comes down to solving all its constituting Sequential Cases.

There are compelling reasons to believe that the combination of you stopping and me driving is, in a sense, right. While this will be explored in greater detail throughout the remainder of this book (since we have yet to determine how to assess actions originating from the initial state), we now understand enough to pinpoint the fundamental flaw in Jackson's reasoning: Once you choose to drive, the situation shifts to a different state and, consequently, to a new decision situation. In this new context, only options for *me* remain, as it has become *my* individual decision situation (your action is now simply part of the context). Naturally, in this context, it is right for me to stop—precisely *because* you are driving, and thus, I find myself in this individual decision situation resulting from the reduction of our original collective decision situation.

However—and this is the crucial point—this reasoning says nothing about which individual options are right in the *initial* state. The distinction introduced by the "because" is significant, as it shifts the frame of reference: I am now in a different decision situation than we were in before. Whatever is right or wrong for me in this new situation does not determine what was right or wrong in the collective decision situation from which we arrived (cf. Figure 7.10). To assume otherwise falls prey to what could be called the collective original sin of consequentialism: the CSM. This misguided attempt to close deontic gaps by incorporating such reasoning led directly into the

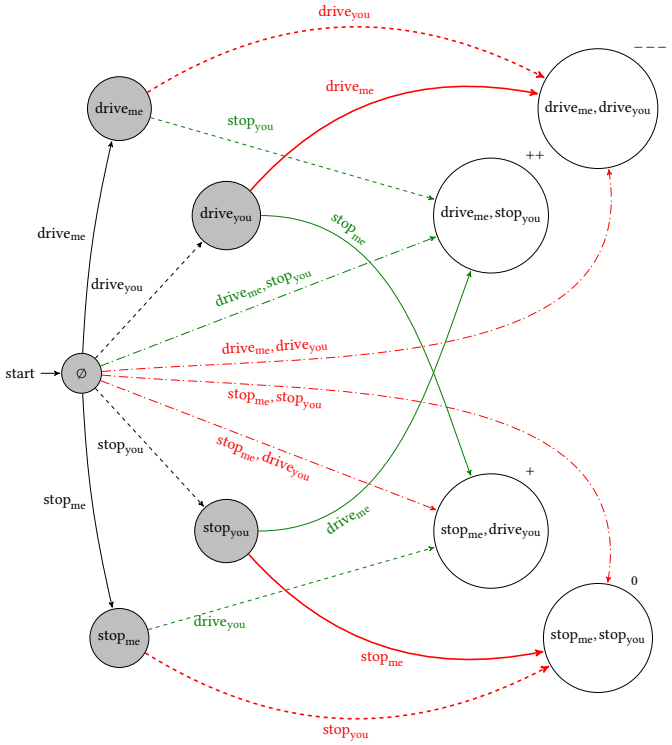


Figure 7.10: The GEF of Intersection with annotations reflecting Jackson's reasoning. The diagram illustrates how the combination of me driving and you stopping is deemed right in the initial state, according to his reasoning. However, once you decide to drive, it becomes right for me to stop (*because* you drive). At this point, another decision situation arises remains, along with a 'new' option. Consequently, it is (the least) highly misleading to claim that the 'right combination' of me driving and you stopping *contains* a wrong action, such as me driving. Rather, the combination is right (in Jackson's terms) relative to the initial state, while me driving is wrong relative to the state labeled "$\text{drive}_{\text{you}}$".

Challenge (cf. Chapter 6). Jackson's observation, therefore, exposes a fundamental flaw, not an asset, in the standard framework of consequentialism, which, until now, has lacked the necessary structural depth to address such cases.

7.3 Filling Gaps With Multi-Agent Amendments

Given the Approach, a solution of the Real Challenge seems within reach. Compositionalism has been rendered defensible, and we have identified an individual decision situation for each agent in each collective decision situation, including proper consequences for each of the agents' options. Camp moac thus merely needs a matching puzzle piece to close the gaps, i.e., a way to assess these options morally so that they fit moac's ambitions. As mentioned before, I call these theoretical extensions to the consequentialist toolbox *multi-agent amendments*.

As it turns out, there are many *prima facie* promising candidates. Unsurprisingly, they all have in common that they ultimately make use of and rely on the already given evaluations of the final outcomes. This makes them consequentialist amendments.

Two classes of amendments can be distinguished. One consists of *aggregative* amendments. These are extensions of the axiological toolbox in a broad sense (cf. Subsection 2.3.3), adding a third dimension of aggregation to the two traditional aggregation dimensions over time and over moral patients. Thus, these methods are 'interpossible', aggregating over how things might possibly unfold. Some of these methods are probabilistic and involve objective probabilities—or, at any rate, can be thus extended. Aggregative amendments allow us to rank the newly discovered outcomes, i.e., the remaining decision situations of the other agents, in such a way that MOCOR can be applied without need for modification.

The other class of amendments is *non-aggregative*. They get by without evaluating the corresponding intermediate outcomes and still arrive at a *ranking of options based on their consequences' moral quality*. Thus, the resulting overall theories remain, in a sense, MOAC theories, even though they might make a modification of MOCOR necessary.

In the following, I present a selection of amendments that I consider to be particularly promising and theoretically well-anchored. Most are based on methods for decision-making under uncertainty or under risk (cf. Subsection 7.1.2). The question of how Camp MOAC should ultimately decide between these amendments will be the subject of the following chapter.

7.3.1 Aggregative Approaches

Aggregative amendments extend the valuation function Val. While Val has so far been defined over final outcomes—and, as a shorthand, directly applied to options[138]—it must now accommodate the intermediate outcomes introduced by the APPROACH. To enable MOCOR to take advantage of these new intermediate outcomes effectively, Val must be capable of evaluating both individual and collective decision situations. (This is particularly important in cases involving more than two agents, where the consequences of the first-acting agent's action still constitute a collective decision situation, albeit involving only the remaining agents.) This approach to closing deontic gaps aligns seamlessly with the consequentialist evaluation pipeline outlined in Section 2.3. Aggregative amendments essentially extend the pipeline by incorporating an evaluation for sets of possible outcomes—an approach familiar from subjective forms of consequentialism (cf. Figure 7.11).

[138]Recall that for an individual decision situation D with actual context C, $\mathsf{Val}_C(\phi)$ for $\phi \in \Phi_D$ was defined as a shorthand for $\mathsf{Val}(\mathsf{Out}_C(\phi))$ (cf. page 36).

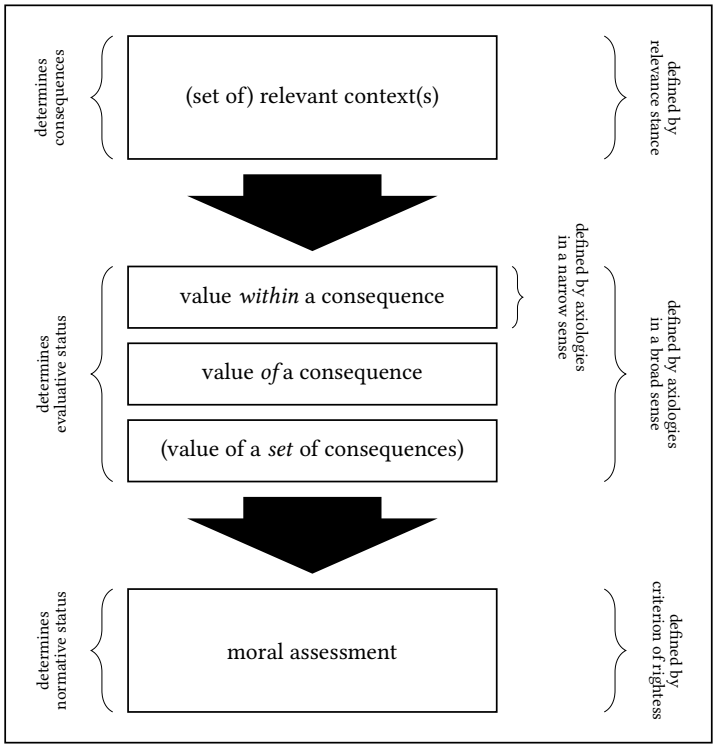


Figure 7.11: How the modules of a consequentialist theory interlink to arrive at moral assessments. Aggregative amendments simply give MOAC theories a way to evaluate sets of consequences, as is common for subjective varieties of consequentialism.

Before introducing the candidates for aggregative amendments, it is helpful to revisit some notational details. First, recall that, with the help of APPROACH, we can decompose arbitrary collective decision situations D with $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$ and actual context C into $n := |\mathcal{A}|$ individual decision situations $\mathsf{D}_{A_1}, \ldots, \mathsf{D}_{A_n}$ with the structure

$$\mathsf{D}_{A_i} := \left\langle A_i,\ \Phi_{A_i},\ \mathsf{Out}_C :\ \Phi_{A_i} \to \mathcal{O}_D^{A_i} \right\rangle$$

where

$$\mathcal{O}_D^{A_i} := \{\, D_{\downarrow\phi} \mid \phi \in \Phi_{A_i} \,\}.$$

Let us call the set

$$\mathfrak{I}_D := \{\, \mathsf{D}_{A_i} \mid A_i \in \mathcal{A}_D \,\}$$

the *decomposition of D*.

Further, recall that we defined $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$, where

$$\mathcal{W} := \bigcup_{D \in \mathfrak{I}} \mathcal{O}_D.$$

(For the sake of simplicity, I assume $\mathcal{V} = \mathbb{R}$ for the rest of this project, although less restrictive constraints would in principle suffice.) Finally, be reminded

that we introduced the short hands $\mathsf{Val}_C(\phi)$ and $\mathsf{Val}_C(\Upsilon)$ for $\mathsf{Val}(\mathsf{Out}_C(\phi))$ and $\mathsf{Val}(\mathsf{Out}_C(\Upsilon))$.

With these in mind, we can now introduce the first amendment.

7.3.1.1 Summation

The two methods of aggregation across moral patients (i.e., individuals or entities considered for their capacity to be morally affected and thus the 'vessels' of moral value) and points in time, typically associated with MOAC, rely on simple summations. It is, therefore, fitting to start with the following straightforward candidate amendment:

Definition 7.2 (Summation) *Let D be an individual decision situation with $\mathsf{D} = \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O} \rangle$ and actual context C. Further, let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be some valuation function. The* value of an individual decision situation D *is*

$$\mathsf{Val}^{\Sigma}(D) := \sum_{\phi \in \Phi} \mathsf{Val}_C(\phi).$$

Let D be a collective decision situation with actual context C and decomposition $\mathrm{I}_D = \{ \mathsf{D}_A \mid A \in \mathcal{A}_D \}$ and let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be some valuation function. The value of a collective decision situation D *is*

$$\mathsf{Val}^{\Sigma}(D) := \sum_{\mathsf{D}_A \in \mathrm{I}_D} \mathsf{Val}^{\Sigma}(\mathsf{D}_A).$$

So, we first lift the valuation function Val Val from the level of outcomes to the level of individual decision situations by defining the value of an individual decision situation as the sum of the values of its possible outcomes, i.e., by summing up all possible outcomes of the situation. Then, we recursively define the value of a collective decision situation as the sum of the values of its individual decompositions. This recursive definition is guaranteed to be well-defined for cases with a finite number of agents (and decisions). This is because the value of each individual decision situation is itself defined as the sum of the values of its possible outcomes, reducing the remaining decision situations further and further until the valuation terminates at the level of final outcomes.

We remember our running example Two Factories. Recall

		Ben	
		pollute	produce cleanly
Ann	pollute	second-worst	worst
	produce cleanly	worst	best

If we assume a few concrete values, we can now directly determine the valuation of Ann's options according to Val^{Σ}:

		Ben		$\mathsf{Val}^{\Sigma}_{C}(\cdot)$
		pollute	produce cleanly	
Ann	pollute	-1000	-2000	-3000
	produce cleanly	-2000	$+1000$	-1000

Thus, according to MOAC with the Summation amendment, it is unconditionally right for Ann in Two Factories to produce cleanly because -1000 is greater than -3000. The same reasoning applies to Ben:

		Ben		$\mathsf{Val}^{\Sigma}_{C}(\cdot)$
		pollute	produce cleanly	
Ann	pollute	-1000	-2000	-3000
	produce cleanly	-2000	$+1000$	-1000
$\mathsf{Val}^{\Sigma}_{C}(\cdot)$		-3000	-1000	

We see that MOAC with Summation 'leads' to the best result in this case.

This is not a contingent finding for the specific values chosen. We can represent Two Factories more generally as

		Ben		$\mathsf{Val}^{\Sigma}_{C}(\cdot)$
		pollute	produce cleanly	
Ann	pollute	v_{pp}	v_{pc}	$v_{pp}+v_{pc}$
	produce cleanly	v_{cp}	v_{cc}	$v_{cp}+v_{cc}$
$\mathsf{Val}^{\Sigma}_{C}(\cdot)$		$v_{pp}+v_{cp}$	$v_{pc}+v_{cc}$	

where $v_{cp}=v_{pc}<v_{pp}<v_{cc}$. Even more generally we have

		Y		$\mathsf{Val}^{\Sigma}_{C}(\cdot)$
		ϕ_Y	$\neg\phi_Y$	
X	ϕ_X	v_{XY}	$v_{X\neg Y}$	$v_{XY}+v_{X\neg Y}$
	$\neg\phi_X$	$v_{\neg XY}$	$v_{\neg X\neg Y}$	$v_{\neg XY}+v_{\neg X\neg Y}$
$\mathsf{Val}^{\Sigma}_{C}(\cdot)$		$v_{XY}+v_{\neg XY}$	$v_{X\neg Y}+v_{\neg X\neg Y}$	

where $v_{X\neg Y}=v_{\neg XY}<v_{XY}<v_{\neg X\neg Y}$.

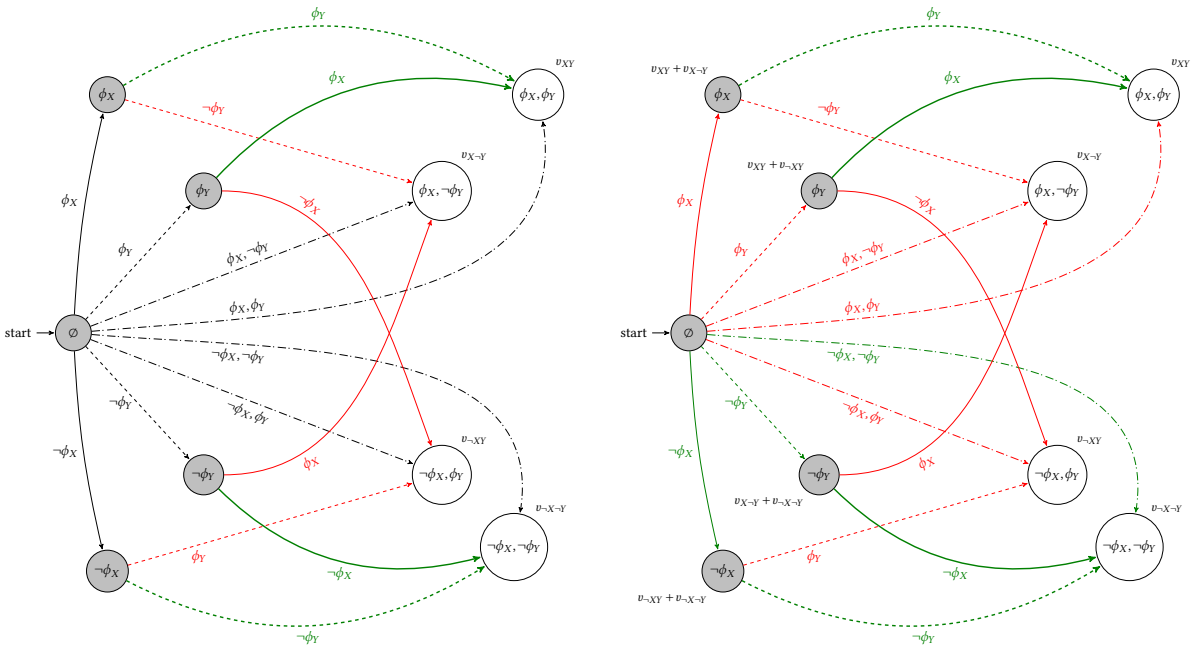


Figure 7.12: The GEF of Two Factories–like cases with value annotations and moral assessment markings (based on MOCOR, under the assumption that $v_{X\neg Y} = v_{\neg XY} < v_{XY} < v_{\neg X\neg Y}$): green edges represent right actions, red edges represent wrong ones, and gray edges correspond to actions for which MOCOR provides no assessments, offering only future-involving conditional evaluations. The left-hand side depicts MOCOR alone, while the right-hand side includes MOCOR with Summation as the amendment of choice (induced values at intermediate states are displayed in bold). Note that the only fully green paths lead to the global optimum.

Thus, we know that $v_{XY} < v_{\neg X\neg Y}$. Since $v_{X\neg Y} = v_{\neg XY}$, we can rewrite this inequality as both $v_{XY} + v_{\neg XY} < v_{X\neg Y} + v_{\neg X\neg Y}$ and $v_{XY} + v_{X\neg Y} < v_{\neg XY} + v_{\neg X\neg Y}$. Therefore, $\mathsf{Val}_C(\phi_z) < \mathsf{Val}_C(\neg\phi_z)$ (for $z \in \{X, Y\}$) independently of the concrete choice of values.

Consequently, MOAC with Summation produces assessments for Troublemakers such as Whiff and Poof and Two Factories that align well with the spirit of PMH. Figure 7.12 illustrates how the gaps have been effectively closed in the annotated GEF of Two Factories.

As a side effect, we can now provide a plausible answer to the question of the borderline case of simultaneous action: we only need to consider the actions of the individual agents and their assessments in the initial state. The indirect assessment of the action combinations arguably follows from this in a manner consistent with Methodological Individualism: only the combination of actions consisting exclusively of right actions is considered right in the derived sense.

I will not carry out analogous considerations in this generality and at this level of detail for the following amendments. Instead, I will restrict myself to evaluating one or two example cases. It stands to reason, however, that we

can learn something from this insight about how camp MOAC should choose among the various amendment candidates, namely by looking at the results of a given choice relative to classes of collective decision situations.

7.3.1.2 Maximization

Next, we consider another type of aggregation as an amendment, which, in a sense, explicitly honors the "M" in MOAC. The basic idea is to approach the task of closing the gap by thinking from the end. Instead of starting one's considerations from the initial state and looking where the first actions might lead, one starts with the values of the final outcomes. One then assesses all actions within some combinations of actions according to the best final outcome they can lead to.

To get to the heart of this idea, we use an auxiliary definition: Let D be a collective decision situation with domain Ψ and actual context C. Further, let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be some valuation function. For an arbitrary $\phi \in \Phi_A \in \Gamma_D$, we define the set of combinations that involve ϕ as

$$\Psi_\phi := \{ \Upsilon \in \Psi \mid \phi \in \Upsilon \}.$$

Based on this set, we can now 'push down' the valuation function from the level of combinations to individual actions.

Definition 7.3 (Maximization$_{\text{PROTO}}$) *Let D be a collective decision situation with actual context C and decomposition $\mathfrak{I}_D := \{ \mathsf{D}_A \mid A \in \mathcal{A}_D \}$ and let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be some valuation function.*

For some A in D and $\phi \in \Phi_A$, the value of ϕ *is*

$$\mathsf{Val}_C^{\max}(\phi) := \max_{\Upsilon \in \Psi_\phi} \mathsf{Val}_C(\Upsilon).$$

This formulation of the idea, notably, does not rely on the APPROACH. As presented, Maximization appears somewhat ad hoc from a consequentialist perspective because, contrary to expectations, the moral assessment of an action is no longer a direct function of the moral quality of *all* its possible consequences. Instead, it depends solely on a particular *possible* consequence, considered in the context of the actions taken by other agents.

With the APPROACH backing us up, however, we can get around this issue. Instead of modifying the valuation function from combinations to options with indeterminate consequences, we can extend it to decision situations, much as we did with Summation:

Definition 7.4 (Maximization) *Let D be an individual decision situation with $\mathsf{D} = \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O} \rangle$ and actual context C and let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$*

be some valuation function. The value of an individual decision situation D *is*

$$\mathsf{Val}^{\max}(D) := \max_{\phi \in \Phi} \mathsf{Val}_C(\phi).$$

Let D be a collective decision situation with actual context C and decomposition $\mathfrak{I}_D = \{ \mathsf{D}_A \mid A \in \mathcal{A}_D \}$ and let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be some valuation function. The value of a collective decision situation D *is*

$$\mathsf{Val}^{\max}(D) := \max_{\mathsf{D}_A \in \mathfrak{I}_D} \mathsf{Val}^{\max}(\mathsf{D}_A).$$

The resulting 'gap fillings' based on these two definitions of Maximization are indeed extensionally equivalent.[139] However, this approach eliminates the need to adapt MOCOR, as we can directly assess the immediate consequences of the respective options using the structure provided by the APPROACH.

For illustration, we can apply this amendment to Troublemakers. This time, we replace the generic notation with a notation of "+" and "−" signs, which makes the order of the values of the final outcomes immediately evident.

		Y		$\mathsf{Val}_C^{\max}(\cdot)$
		ϕ_Y	$\neg\phi_Y$	
X	ϕ_X	−	−−−	−
	$\neg\phi_X$	−−−	+++	+++
$\mathsf{Val}_C^{\max}(\cdot)$		−	+++	

We thus see that MOAC together with Maximization is again, like it was in combination with Summation, guaranteed to lead to the best possible result for all Troublemakers with this valuative profile. As with Summation, this is no contingent finding, since it is independent of the specific values of the final outcomes, but it is implied by the order of these values.

[139] The proof is left to the interested reader, but the intuition is as follows: The value of an individual action is equal to the value of its consequence, consistent with the shorthand notation adopted earlier. According to the APPROACH, the consequence of an individual agent's action is the reduced collective decision situation relative to that action. If the reduced situation is an individual decision situation, then, per the first part of Definition 7.4, its value is the value of the best final outcome(s) reachable from it. If the reduced situation remains collective, then, per the second part of Definition 7.4, its value is the value of the best decision situation reachable from it. Ultimately, because we are dealing with a finite number of agents, we will eventually arrive at the former (individual decision situation) case. Thus, the final result is that the value of each option in the original collective situation corresponds to the value of the best final outcomes reachable by any combination that includes that option, which aligns precisely with Definition 7.3.

7.3.1.3 Expected Utility

The following amendment, inspired by decision-theoretic (or prospective) Maximizing Subjective Act-Consequentialism, seeks to make the concept of expected value applicable to objective consequentialism. This approach can also be seen as an application (or adaptation) of Kagan's proposed solution to MOAC theories (cf. Subsection 4.4.1). The primary *theoretical* drawback of this approach is its reliance on objective probabilities—both concerning the actions each agent will take and the sequence in which they will act. For one thing, objective probabilities are metaphysically dubious. For another, even setting aside these metaphysical concerns, it remains entirely unclear where to obtain appropriate probability distributions. Furthermore, as we will see, this approach also faces an *extensional* drawback. For these reasons, there are compelling arguments against MOAC adopting EU.

The advantage of this approach lies in its ability to integrate the actual or expected actions of other agents into the moral assessment of an agent's options using a well-established and familiar framework, allowing MOAC to make conceptual space for a rigorous and well-understood treatment of risks. Consider a scenario in which Ann, based on numerous interactions with Ben, knows that he is likely to pollute. Suppose, for example, that Ben has a track record of environmentally harmful actions, providing strong evidence of his bad disposition. Within an approach grounded in expected values, it is clear how Ann's experiences with Ben would factor into her moral deliberations. In contrast, we must ask whether an extension of MOAC to multi-agent scenarios should instead consistently base its assessments on the assumption that all agents act rightly while dismissing facts, such as Ben's assumed disposition to damage the environment, in favor of purely theoretical assumptions. If so, being a solitary 'good agent' among persistently bad agents would often lead to severe, yet otherwise foreseeable and thus avoidable, moral disasters. Incorporating risk awareness into MOAC is supported by the same considerations that make it appealing in subjective settings: a pragmatic balance between aiming for optimal outcomes and accounting for uncertainty or risk in decision-making.

I am by no means the first to have this idea. In the essay already quoted in the last chapter, A. N. Prior (Prior and Raphael 1956, p. 93) writes, basically in response to the REAL CHALLENGE:

> Taking the non-determinist horn first, perhaps we can say that if determinism is not true, it suffices to speak of a duty to do what will *probably* have the best total consequences of all the actions open to us. We can only take this line, however, if we are prepared to talk about objective probabilities; that is, if we are prepared to argue that "p is probable" need not merely mean "We don't know that p will be true, but what evidence we have is more in favour of it than against it", but may mean something more like "p is not

> yet either going to be the case or not going to be the case, but is more like going to be the case than not".

I think that the burden of objective probabilities brings too much theoretical ballast for MOAC. While I do not want to exclude that consequentialists find a way to carry this burden and make it fruitful in the practice of normative ethics, I do not make it my project to defend this approach here. So, I merely sketch here what the corresponding amendment would look like and then shelve it.

As previously mentioned, for each individual decision situation, we require a probability distribution over the possible actions of all agents. Similarly, for each collective decision situation, a probability distribution over the possible temporal orderings of the agents' actions is necessary.[140] In the following, I use $\Pr_X$ to denote a given probability distribution over a set X. (Later, I will also use $\Pr(X)$ to refer to the set of all possible probability distributions over X, such that $\Pr_X \in \Pr(X)$.) With this notation in place, we can now define:

Definition 7.5 (Expected Utility) *Let D be an individual decision situation with $\mathsf{D} = \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O} \rangle$ with actual context C and let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be some valuation function. The* (expected) value of an individual decision situation *D is*

$$\mathsf{Val}^{\mathrm{EV}}(D) := \sum_{\phi \in \Phi} \Pr_\Phi(\phi) \mathsf{Val}_C(\phi).$$

Let D be a collective decision situation with actual context C and decomposition $\mathbb{I}_D := \{ \mathsf{D}_A \mid A \in \mathcal{A}_D \}$ and let $\mathsf{Val}_C : \mathcal{W} \to \mathbb{R}$ be some valuation function. The (expected) value of a collective decision situations *D is*

$$\mathsf{Val}^{\mathrm{EV}}(D) := \sum_{\mathsf{D}_A \in \mathbb{I}_D} \Pr_{\mathcal{A}}(A \text{ acts next}) \cdot \mathsf{Val}^{\mathrm{EV}}(\mathsf{D}_A).$$

The (expected) value of an option ϕ within a collective decision situation *D is*

$$\mathsf{Val}^{\mathrm{EV}}_C(\phi) := \mathsf{Val}^{\mathrm{EV}}(D_{\downarrow\phi}).$$

Since I don't explore Expected Utility further in this book, I don't go into much detail about what this amendment implies with respect to Troublemakers. However, it is worth looking briefly at the following instance of Two Factories:

[140] Notably, this second probability distribution need not be of a strictly metaphysical nature. For example, we might propose that every sequence of actions is equally probable, potentially treating simultaneous actions as quasi-impossible at a sufficiently fine-grained 'temporal resolution'. However, even this assumption demands justification, which cannot be provided here, particularly within the context of an Expected Utility–focused narrative. (Such a more analytical approach will be revisited in a different context in the next chapter.)

			Ben		$\mathsf{Val}^{\mathrm{EV}}_C(\cdot)$
	$\downarrow \Pr_{\Phi_{\text{Ann}}}(\cdot)$	$\Pr_{\Phi_{\text{Ben}}}(\cdot) \rightarrow$	0.8 pollute	0.2 produce cleanly	
Ann	0.8	pollute	−1000	−2000	−1200
	0.2	produce cleanly	−2000	+1000	−1400
$\mathsf{Val}^{\mathrm{EV}}_C(\cdot)$			−1200	−1400	

What this instance shows is that the result of following $\mathsf{Val}^{\mathrm{EV}}$ *stringently* can be suboptimal (because, given the chosen probability distributions, producing in a polluting manner has a higher expected value for both agents than producing clean so both would act rightly if they produced dirty). Thus, in essence, filling the deontic gaps using $\mathsf{Val}^{\mathrm{EV}}$ apparently provides room for the Challenge to re-emerge.

What we also observe is that the non-conditional assessments of $\mathsf{Val}^{\mathrm{EV}}$-amended MOCOR depend critically on the specific probability distributions involved.[141] In other words, traditional descriptions of collective decision situations are strongly incomplete relative to this approach (unsurprising, as objective probabilities are rarely considered in the context of MOAC). Furthermore, our definition of symmetry would require significant revision to incorporate probability distributions alongside the valuative profile. In conclusion, while Expected Utility might hold some potential as a position within camp MOAC, adopting it introduces substantial theoretical complexities and complications. At the same time, there is no guarantee that it would effectively address the Challenge. Consequently, I set aside Expected Utility.

Of course, there are a myriad of other ways to evaluate in an aggregative manner the consequences newly gained thanks to the Approach. However, I believe that the amendments presented here are the most important ones that deserve closer consideration, given established methods from decision and game theory and in light of other consequentialist defaults. So, let us turn to non-aggregative amendments next.

7.3.2 Non-Aggregative Amendments

Non-aggregative amendments enable the ranking of agents' options in collective decision situations based on consequentialist principles—that is, relying solely on the moral quality of the options' immediate consequences as provided by the Approach, but *without* directly evaluating these consequences. The advantage of such amendments is that many of them can operate independently of the Approach and, therefore, do not necessitate

[141] For instance, the assessments above would already differ for distributions such as $\Pr_{\Phi_A}(\text{pollute}) = 0.75$ and $\Pr_{\Phi_A}(\text{produce cleanly}) = 0.25$ for both agents, i.e., $A \in$ Ann, Ben.

a broader extension of axiology. However, the disadvantage lies in a challenge we have already encountered with Maximization: for these amendments to gain traction, MOCOR must be adapted. Additionally, without leveraging the APPROACH, these adaptations often appear *ad hoc*, as the connection between the evaluated option and its specific consequences becomes somewhat obscured.

In the following, I introduce five amendments from the many conceivable ones, which I consider the most straightforward and promising. These amendments fall into three distinct groups.

7.3.2.1 (Non-)Domination

Based on the notion of Ψ_ϕ from above, we can define what it means that one combination dominates the other. For two options $\phi, \phi' \in \Phi_A \in \Gamma_D$ of the same agent, we write $\phi \geq \phi'$ for "ϕ weakly dominates ϕ'" and write $\phi > \phi'$ for that "ϕ strongly dominates ϕ'". We define the following, where C represents a variable for a given context:[142]

$$\phi \geq \phi' \quad \text{if and only if} \quad \forall\, \Upsilon \in \Psi_\phi, \forall\, \Upsilon' \in \Psi_{\phi'} : \mathsf{Val}_C(\Upsilon) \geq \mathsf{Val}_C(\Upsilon')$$

and, accordingly,

$$\phi > \phi' \quad \text{if and only if} \quad \forall\, \Upsilon \in \Psi_\phi, \forall\, \Upsilon' \in \Psi_{\phi'} : \mathsf{Val}_C(\Upsilon) > \mathsf{Val}_C(\Upsilon').$$

This allows us to define two[143] collective criteria of rightness:

Definition 7.6 (Dominance) *Let D be a collective decision situation with $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$, where $A \in \mathcal{A}$ is some agent, $\phi \in \Phi_A \in \Gamma$ is an option of that agent A, and C is the actual context. Additionally, let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be a valuation function. Option ϕ of some A in D is* right for *A in C if and only if ϕ weakly dominates every other option of A in C, i.e., if and only if $\forall \phi' \in \Phi_A : \phi \geq \phi'$.*

Definition 7.7 (Non-Dominated) *Let D be a collective decision situation with $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$, where $A \in \mathcal{A}$ is some agent, $\phi \in \Phi_A \in \Gamma$*

[142] Note that these relations implicitly depend on a valuation function Val. To avoid overloading the notation with indices, I have chosen not to explicitly indicate this dependency, although the valuation function is explicitly stated in the definitions below.

[143] Non-Dominated corresponds to what John F. Horty (2001, chapter 4) once called "dominant act utilitarianism". To the best of my knowledge, Horty's purely formal work on multimodal deontic logic, limited to stit semantics, is the closest thing to my project. My thoughts have developed independently of Horty's book, which I came across only later through a direct conversation with him at a Dagstuhl seminar in 2016. I am deeply indebted to Jeff and his work, though.

is an option of that agent A, and C is the actual context. Additionally, let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ *be a valuation function. Option* ϕ *for A* right for *A in C if and only if* ϕ *is not strongly dominated by any other option of A in C, i.e., if and only if* $\nexists \phi' \in \Phi_A : \phi' > \phi$.

Both of these criteria offer distinct moral guidance in collective scenarios. Reconsider our standard example of a Troublemaker:

		Y			
		ϕ_Y	$\neg\phi_Y$	is dominant	is non-dominated
X	ϕ_X	$-$	$---$	false	true
	$\neg\phi_X$	$---$	$+++$	false	true
is dominant		false	false		
is non-dominated		true	true		

This application of the two criteria reveals their limitations. While Dominance corresponds to a well-known stringent standard for decision-making, it is rare for a single option to dominate all others. Such a criterion can often be overly restrictive. True, adopting Dominance would allow MOAC to close the moral gaps. However, agents might frequently find themselves in scenarios without any morally right option. Accepting Dominance, therefore, comes with the significant cost of embracing numerous true moral dilemmas, forcing this version of MOAC to abandon any aspirations of achieving No Moral Dilemmas. Consequently, it is unrealistic to expect such a version to satisfy PMH in any meaningful sense. Moreover, when a dominant option *does* exist, Conditionalization combined with the Sure-Thing Principle will necessarily yield the same result as combining MOCOR with Dominance. For these reasons, Dominance can be set aside for the remainder of this project.

In contrast, Non-Dominated ventures to the opposite extreme. This criterion implies a high degree of permissiveness. In interesting collective decision scenarios, which Troublemakers are by design, multiple options will not be dominated by any others. As a result, under the Non-Dominated criterion, a large number of actions may simultaneously be considered right, undermining the ability to provide the specific and actionable moral guidance that PMH demands.

Fortunately, dominance-based criteria are not the only non-aggregative amendments available. The field of decision-making under uncertainty offers numerous well-established methods, three of which are particularly worth introducing in the context of this project.

7.3.2.2 MaxiMin and MaxiMax

Whoever mentions "dominance" has to mention "MaxiMin" and "MaxiMax" as well. These twin principles, foundational in decision and game theory, are often discussed in tandem due to their contrasting orientations toward risk. As we have seen, dominance-based amendments, in a sense, zero in on universally preferable options. On the other hand, when dominance doesn't provide clear guidance, which is often the case and certainly is so with Troublemakers (as shown above), MaxiMin and MaxiMax (and a zoo of further criteria) can step in. They spotlight the strategic distinction between safeguarding against worst-case scenarios and gunning for the best possible outcomes. Beyond theoretical significance, the MaxiMin principle, in particular, has also gained a reputation in terms of ethical theorizing, primarily, of course, because of its place value in John Rawls' (1971) theory of justice. Very roughly, Rawls advocated for a society where inequalities are treated such that the least advantaged benefit the most, hence advocating a MaxiMin approach in the realm of distributive justice.

As always, we start with rigorous definitions of the corresponding modified versions of MOCOR:

Definition 7.8 (MaxiMin) *Let D be a collective decision situation with $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$, where $A \in \mathcal{A}$ is some agent, $\phi \in \Phi_A \in \Gamma$ is an option of that agent A, and C is the actual context. Additionally, let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be a valuation function. Option ϕ is* right for *A in C if and only if the worst final outcome associated with ϕ, given C, is at least as good as the worst final outcome associated with any other option ϕ' available to A in light of C. Formally:*

$$\forall \phi' \in \Phi_A : \quad \min_{\Upsilon' \in \Psi_{\phi'}} \mathsf{Val}_C(\Upsilon') \leq \min_{\Upsilon \in \Psi_\phi} \mathsf{Val}_C(\Upsilon).$$

Definition 7.9 (MaxiMax) *Let D be a collective decision situation with $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$, where $A \in \mathcal{A}$ is some agent, $\phi \in \Phi_A \in \Gamma$ is an option of that agent A, and C is the actual context. Additionally, let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be a valuation function. Option ϕ is* right for *A in C if and only if the best possible final outcome associated with ϕ, given C, is at least as good as the best possible final outcome associated with any other option ϕ' available to A in light of C. Formally:*

$$\forall \phi' \in \Phi_A : \quad \max_{\Upsilon' \in \Psi_{\phi'}} \mathsf{Val}_C(\Upsilon') \leq \max_{\Upsilon \in \Psi_\phi} \mathsf{Val}_C(\Upsilon).$$

Here is, once again, an application of these two amendments to a Troublemaker with the default structure:

		Y		min	max
		ϕ_Y	$\neg\phi_Y$		
$\prec$	ϕ_X	$-$	$---$	$---$	$-$
	$\neg\phi_X$	$---$	$+++$	$---$	$+++$
min		$---$	$---$		
max		$-$	$+++$		

According to MaxiMin, every option is deemed right, whereas MaxiMax identifies only $\neg\phi$ as right. Notably, MaxiMax turns out to be extensionally equivalent to Maximization. MaxiMin, in contrast, is highly permissive and generally fails to guarantee the best outcome if all agents act in accordance with its assessments.[144]

Before we can turn to the final evaluation of amendments, there remains one variant that is worth introducing.

7.3.2.3 Mixed Strategies

There is another method that, again, requires us to enter the domain of probabilistic decision-making: *mixed strategies*. Mixed strategies provide a probabilistic approach to decision-making that introduces flexibility to the evaluation of options. While it involves objective probabilities, it comes without metaphysical baggage as these probabilities are analytic, i.e., they are derived. mathematically from the structure of the case at issue. The central idea is this: Instead of committing to a single deterministic action, agents adopt probability distributions over their available options, allowing them to optimize the expected outcome. This effectively expands the option space to include infinitely many new options, as agents can choose actions with countless different probabilities. While this may seem to complicate decision-making, it offers a balanced framework that accommodates both risk and adaptability.

At first glance, mixed strategies might appear to unjustifiably modify collective decision situations by enriching the option spaces. However, this expansion has several compelling advantages. First, it aligns with the intuitive observation that decision-making is not always deterministic. Reflecting on real-life scenarios, it seems plausible (at least to me) that if faced with the same situation multiple times, one might not always make the *same* choice—even with identical epistemic, emotional, and contextual states. This variability

[144]Both criteria can be refined by introducing their lexicographic versions. For instance, MaxiMin (MaxiMax) can be modified to rank options with equally good worst (best) outcomes by considering the second worst (best) outcome, then the third, and so on, recursively, until only options with identical evaluative profiles are ranked equally. The details of these refinements are omitted here but will be briefly revisited later.

suggests that acting according to probability distributions over one's options is not only possible but might, in some cases, be morally or rationally appropriate. Mixed strategies, therefore, might very well capture a nuanced aspect of human agency that purely deterministic models overlook.

Additionally, mixed strategies have proven to be theoretically fruitful in numerous fields, particularly in game theory, where they enable rigorous equilibrium analysis. They often reveal Nash equilibria[145] where purely non-probabilistic approaches fail to identify any optimal strategy. In other words, by leveraging the tools of probability and expected utility theory, mixed strategies enable agents to optimize outcomes in ways that wouldn't be possible with a limited set of *pure* strategies. In essence, while the approach might seem counterintuitive, it offers a richer, more reflective, and more analytically powerful framework for understanding decision-making in complex scenarios.

Moreover, concerns about the formal feasibility of expanding option spaces to infinity are unfounded. The mathematical toolbox required to handle such scenarios is well-developed and fully capable of accommodating these complexities. While introducing mixed strategies entails certain theoretical burdens, these are significantly lighter than the challenges posed by Expected Utility. Importantly, the probabilities involved in mixed strategies are analytical constructs rather than metaphysical claims, which makes them more tractable and less conceptually contentious.

In summary, while mixed strategies may initially seem counterintuitive, they offer a powerful and reflective framework for understanding decision-making under uncertainty that comes naturally with multi-agent scenarios.

Beyond all that, the idea of incorporating mixed strategies is far from new. In fact, it has a long history. In this context, it is worth revisiting the influential double essay on utilitarianism by Bernard Williams and J. C. C. Smart (1973). While we have already discussed Bernard Williams's (1973) example of Job Market—later employed by Jonathan Glover (Glover and Scott-Taggart 1975) to spark the debate on the Challenge (cf. Section 3.4, page 53)—J. C. C. Smart's (1975) response to Williams sometimes goes unnoticed. Specifically, his intriguing proposal to extend moac into collective contexts through the introduction of mixed strategies, presented as a response to Richard Brandt's Frenchman case (cf. Section 3.4), deserves closer attention. Recall the example case, previously introduced on page 63 and repeated here for convenience:

> Suppose that, in wartime England, people are requested, as a measure essential for the war effort, to conserve electricity and gas by having a maximum

[145] A Nash equilibrium, named after John Nash (1950; 1951), represents an outcome in a game (i.e., a collective decision situation) where no player can unilaterally improve their outcome given the strategies of others. These equilibria correspond to stable states of mutual best responses. In a sense, we encountered such equilibria earlier: non-global equilibria are what made Troublemakers challenging (cf. Subsection 6.4.1).

> temperature of 50 degrees F. in their homes. A utilitarian Frenchman living in England at the time, however, argues as follows: "All the good moral British obviously will pay scrupulous attention to conforming with this request. The war effort is sure not to suffer from a shortage of electricity and gas. Now, it will make no difference to the war effort whether I personally use a bit more gas, but it will make a great deal of difference to my comfort. So, since the public welfare will be maximized by my using gas to keep the temperature up to 70 degrees F. in my home, it is my duty to use the gas."
>
> According to the act-utilitarian theory, this argument is perfectly valid. But we should not take it seriously in fact. Why not? At least part of the reason is that we think that, if a sacrifice has to be made for the public good, all should share in it equally. Imagine the outcry in Britain, if it became known that members of the Cabinet, who knew that electricity and gas were in good supply because of the country's willingness to sacrifice, used this argument to justify using whatever power was necessary to keep their homes comfortable.

Smart's response to the challenge was to advocate for mixed strategies (Smart 1973, p. 59):

> There is a circularity in the situation which cries out for the technique of game theory.
>
> There are three types of possibility: (a) he can decide to obey the government's request; (b) he can decide not to obey the government's request; (c) he can decide to give himself a certain probability of not obeying the government's request, e.g. by deciding to throw dice and disobey the government's request if and only if he got a certain number of successive sixes.
>
> To decide to do something of type (c) is to adopt what in game: theory is called "a mixed strategy". On plausible assumptions it would turn out that the best result would be attained if each member of the act-utilitarian society were to give himself a very small probability p of disobeying the government's request.

I think this idea is plausible and worth spelling out in a bit more detail than Smart did. For this, we first define, given a collective decision situation D with actual context C, domain Ψ and some valuation function $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$:

$$\mathrm{EV}(\mathrm{Pr}_\Psi) := \sum_{\Upsilon \in \Psi} \mathrm{Pr}_\Psi(\Upsilon)\mathsf{Val}_C(\Upsilon).$$

What we seek, then, is a probability distribution over the domain Ψ of some collective decision situation, $\mathrm{Pr}_\Psi \in \mathrm{Pr}(\Psi)$, that maximizes $\mathrm{EV}(\mathrm{Pr}_\Psi)$. First, note that Pr_Ψ is simply the joint probability over the actions of the agents involved in the combination, i.e., for a Υ of n actions $\phi_1, \ldots, \phi_n$ we have:

$$\mathrm{Pr}_\Psi(\Upsilon) = \mathrm{Pr}_\Psi\big(\langle \phi_1, \ldots, \phi_n \rangle\big) := \mathrm{Pr}_\Gamma(\phi_1, \ldots, \phi_n).$$

Second, since I have narrowed down the investigation to maximal Coordination Cases, a strong form of Independency of Action is guaranteed (cf. Subsection 5.2.2), $\Pr_\Gamma(\phi_1,\dots,\phi_n)$ can be simply decomposed into a factorization, i.e., into a product of the marginal probability distributions. Thus, given a probability distribution $\Pr_{\Phi_A} \in \Pr(\Phi_A)$ for each agent $A \in \mathcal{A}_D$, we have:[146]

$$\Pr_\Psi(\phi_1,\dots,\phi_n) = \prod_{A\in\mathcal{A}} \Pr_{\Phi_A}(\phi_i).$$

Now we can define:

Principle 7.1 (Mixed Strategies) *Let D be a collective decision situation with $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O}\rangle$, where $A \in \mathcal{A}$ is an agent, $\phi \in \Phi_A \in \Gamma$ is an option of that agent A, and C is the actual context. Additionally, let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be a valuation function. To perform an action $\phi \in \Phi_A$ in D with probability $\Pr_{\Phi_A}(\phi)$ is* right for *A in C if and only if A's choice follows a probability distribution $\Pr_{\Phi_A}$ that belongs to a set of acceptable probability distributions for all agents, such that $\Pr_\Psi$, given by*

$$\Pr_\Psi(\Upsilon) := \prod_{A\in\mathcal{A}} \Pr_{\Phi_A}(\phi_i),$$

maximizes $\mathrm{EV}(\Pr_\Psi)$ relative to C.

It is easy to calculate the searched-for $\Pr_\Psi$ since, by definition,

$$\mathrm{EV}(\Pr_\Psi) = \sum_{\Upsilon\in\Psi} \Pr_\Psi(\Upsilon)\mathsf{Val}_C(\Upsilon).$$

Thus, we can just calculate the distribution $\Pr_\Psi$ such that the derivative vanishes, i.e., $\frac{\partial \mathrm{EV}(\Pr_\Psi)}{\partial \Pr_\Psi} = 0$. In consideration of the function values 'at the boundary' (i.e., for performing one option with a probability of 1), we can thus identify the distributions $\Pr_\Psi$ that maximizes $\mathrm{EV}(\cdot)$, i.e., we can calculate the set

$$\operatorname*{arg\,max}_{\Pr_\Psi\in\Pr(\Psi)} \mathrm{EV}(\Pr_\Psi).$$

Note that, as with the previously discussed non-aggregative amendments, this approach also necessitates modifying MOCOR to incorporate expectation

[146] Note that we can easily translate this notion such that it does not maximality, but only Independency of Action. For this, let us agree on the notation that Υ_A denotes the option $\phi \in \Phi_A$ with $\phi \in \Upsilon$ and define $\mathcal{A}_\Upsilon$ to denote the subset of $\mathcal{A}$ containing all and only those agents that contribute to Υ (and thus Υ_A is defined for every $A \in \mathcal{A}_\Upsilon$ for arbitrary $\Upsilon \in \Psi$ by constructions). Now we can rewrite the factorization to

$$\Pr_\Psi(\Upsilon) := \prod_{A\in\mathcal{A}_\Upsilon} \Pr_{\Phi_A}(\Upsilon_A).$$

Amendment	Remarks	Still in the Game?
Summation	promising	in
Maximization	promising	in
Expected Utility	Although in principle it is not impossible to do something with it, the assumption of objective probabilities brings too great a burden with it. It might be a bigger project to fill this approach with life. Put aside for the time being.	out
Dominance	Allows moral dilemmas and otherwise offers nothing that we could not already derive with the Sure-Thing Principle.	out
Non-Dominated	Too permissive, especially in combination with PMH-driven, optima-demanding intuitions (and this is where we are coming from).	out
MaxiMin	Too permissive in combination with PMH-driven, optima-demanding intuitions.	out
MaxiMax	Extensionally equivalent to Maximization but comes with modification of MOCOR.	out
Mixed Strategies	promising	in

Table 7.1: The introduced amendments, which of them are still in the game, and which are no longer (and, in brief, why the amendments that were thrown out.

values. While this adjustment is not a definitive objection, it does require proper justification—a matter to which I will return in the next chapter. This brings us to the final question: what remains to be addressed?

7.4 What Remains to Be Done

Up to now, we've achieved several insights. First, we have seen how the APPROACH brought genuine consequences in the form of reduced decision situations of individual options to the consequentialist workbench, even in the case of the most complex collective decision situations. Second, we have observed that various amendments to MOAC would, in principle, allow consequentialists to employ these newly discovered consequences in order to fill the gaps that create the REAL CHALLENGE. While several of these amendments could be set aside due to various considerations, others remain viable contenders for addressing the REAL CHALLENGE in an overall promising way. An overview can be found in Table 7.1.

But which of these amendments should MOAC adopt to address not only the REAL CHALLENGE, but also at the same time the CHALLENGE? After

all, we do not just want to fill the gaps, but we want to do so in a way that does not allow the CHALLENGE to reemerge. In light of the Pyramid, it seems natural to focus on the INTERNAL CHALLENGE first and revisit PMH, hoping that it defines MOAC's true collective objective. This would allow us to identify those states in a general extensive form that represent the final outcomes to which MOAC should lead the way so that, in turn, we can then select an amendment that closes the gaps accordingly. So, before we decide on an amendment, it's time to explore the legitimacy and limits of PMH.

Chapter 8

On Reasonable Disharmonies and the Quest for the Best Amendment

At this point, we have a clear picture before us: We have established good reasons to believe that collective decision situations, in the spirit of Compositionalism, are nothing more than a collection of individual decision situations. The consequences that are necessary for this to be possible, long neglected by Camp MOAC, have been provided by the APPROACH. In addition, we have found several theoretical supplements, which I called "amendments", which enable the champions of MOAC to close the gaps causing the REAL CHALLENGE. We were able to reject some of the amendment candidates at this stage. The remaining candidates will be tested for suitability in this chapter.

The question, of course, is what these tests could look like and how, in the end, we should decide between the different amendments. The answer obviously arises from the larger context of this project: we need a solution to the REAL CHALLENGE, and one that also solves the CHALLENGE in its various guises.

In light of the Pyramid, it seems logical to prioritize addressing the INTERNAL CHALLENGE first and then revisit the Principle of Moral Harmony (or shorter: the PMH), with the expectation that it outlines MOAC's 'true' collective objective. This approach would enable us to identify the paths through the generalized extensive form (GEF) that lead to the final outcomes MOAC is intended to guide us toward. In turn, this would inform the selection of an amendment designed to close the gaps appropriately. Therefore, before settling on an amendment, it is essential to examine the legitimacy and limitations of PMH.

The task of this chapter is threefold. First, we revisit PMH and examine its potential guidance on this question. We will quickly realize that MH and the other traditional formulations of PMH are overly stringent, as they conflict

with more foundational objective-consequentialist commitments. By the end of this investigation, we will have arrived at a suitably nuanced yet somewhat vague version of Moral Harmony (MH).

The second objective of this chapter is to refine and elaborate on this revised formulation. To achieve this, I introduce the well-established concept of *policies* into the consequentialist framework and explore how such policies can be evaluated from MOAC's perspective, particularly in the context of our revised version of MH.

Finally, this refined understanding enables us to revisit and assess the previously discussed amendments. Since MOAC, in combination with any amendment, induces a policy, we can now evaluate these policies against the updated framework. By the end of the chapter, I will advocate for a specific amendment and demonstrate how this resolves the challenges posed by the Pyramid in a decisive and unified manner.

8.1 Revisiting Principle of Moral Harmony

In the introduction, I introduced the CHALLENGE by appealing to the principle of:

View 1.1 (Congruence) *The right and the best are congruent in the sense that doing what is right goes (necessarily) hand in hand with bringing about the morally best consequences that can be brought about.*

Building on this fundamental tenet of MOAC, I outlined two readings—-one individual and the other collective. The individual reading asserts that the right action is the one that brings about the best possible consequences. This notion is so central to MOAC theories that it is encapsulated in their defining criterion of rightness:

Principle 2.2 (MOCOR) *Let D be an individual decision situation involving an agent A with a set of options Φ and with actual context C. An action $\phi \in \Phi$ is right for A in D given C if and only if, relative to C, there is no alternative action $\phi' \in \Phi$ with better consequences than ϕ.*

The collective reading, which I referred to as Principle of Moral Harmony (or shorter PMH), was distilled from multiple sources into the following formulation:

Principle 5.1 (Collectively Maximizing) *Let D be a collective decision situation with domain Ψ and with actual context C. If $\Upsilon \in \Psi$ consists only of right actions, then there is (and can be) no alternative $\Upsilon' \in \Psi$ with better consequences than Υ relative to C.*

When we broaden our perspective beyond the typical Troublemakers, it becomes evident, however, that there is a fundamental issue with this second principle. As plausible as Collectively Maximizing may initially appear, a closer examination reveals that it *cannot* hold true—at least not in such generality. In the following discussion, I will first explicate the issue inherent in PMH and argue that the necessary revisions and limitations required by this realization are far more significant than is often tacitly assumed. Subsequently, I will propose a modified formulation of MH, which, while underdetermined in certain respects, serves as an excellent benchmark for guiding our remaining gap-filling efforts.

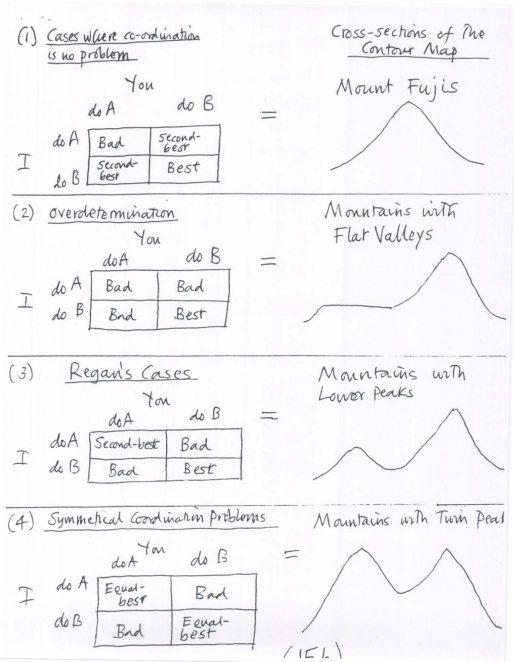


Figure 8.1: Parfit's contour maps (again).

8.1.1 The Limits of PMH

Recall Parfit's contour maps from the previous chapter (see Figure 8.1), which offer a taxonomy of Coordination Cases. Let us take a closer look at what Parfit refers to as "Symmetrical Coordination Cases".[147] To illustrate, consider a modified version of Regan's Whiff and Poof, presented here in its normal form, as encountered multiple times throughout this book:

		Poof	
		not-push	push
Whiff	not-push	10	0
	push	0	10

In such a case, expecting morality to be able to guide the two agents toward one of the best possible outcomes with certainty is tantamount to believing in magic. The combination of symmetry and the non-uniqueness of the optimum poses a significant challenge to such aspirations. To better understand the issue, consider the following reasoning. Suppose T is a moral theory that satisfies Collectively Maximizing. Then it must deem either (1) both Poof

[147] I think this label is a misnomer, as all the cases Parfit includes in this list are symmetrical (cf. Subsubsection 5.2.1.3).

and Whiff pushing their buttons as right and not pushing as wrong, or (2) both Poof and Whiff not pushing their buttons as right and pushing as wrong. Any other assessment would allow for combinations of right actions with suboptimal outcomes—scenarios where one agent pushes while the other does not. But what could possibly be the morally relevant difference between these two moral assessments? Even without invoking Approach, it seems impossible, from a consequentialist standpoint, to identify a morally relevant distinction between the two cases. The rows and columns are merely inverted, yet identical with respect to their valuative profile. For T to satisfy Collectively Maximizing, it would need to arbitrarily single out one optimum and then declare the actions that produce it as right. But why prioritize one optimum over the other when they are indistinguishable in moral terms?

However, the problem goes deeper than merely being an implausible requirement for cases with multiple global optima, because we can construct cases in which satisfaction of Collectively Maximizing is not only implausible but would even be *logically incompatible with objective consequentialism*. Consider the following case, illustrated in Figure 8.2:

Case 8.1 (Seaman Clumsy) *Amid the vast expanse of the high seas, three sailors are aboard a ship: the navigator Mikel, the steerswoman Laika, and a seaman known as Clumsy. As the ship sails through a violent hailstorm, Clumsy, who is a notoriously poor swimmer, suddenly topples overboard.*

Both Mikel and Laika, positioned on opposite ends of the vessel, witness Clumsy's plight. The storm's intensity makes it impossible for them to ascertain whether the other has seen the incident, and the raging elements prevent any form of communication or coordination. Each faces an immediate decision: they can either dive into the turbulent waters to aid Clumsy or turn around to retrieve the lifebuoy to toss to him.

The ship's high bulwark presents a significant challenge. If both Mikel and Laika decide to jump, they will find it impossible to hoist themselves back aboard, dooming all three sailors to the depths. Conversely, if both decide to fetch the lifebuoy, the time Clumsy is left alone in the stormy sea will be too long, and he will succumb to the waves before they can assist.

The optimal scenario, no doubt, would be for one of them to dive in to help Clumsy stay afloat while the other retrieves the lifebuoy. However, if neither of them jumps, Clumsy dies; and if both jump, they all drown. Thus, we may represent Seaman Clumsy in the following normal form:

		Mikel	
		jump	get the buoy
Laika	jump	worst	equal-best
	get the buoy	equal-best	second-worst

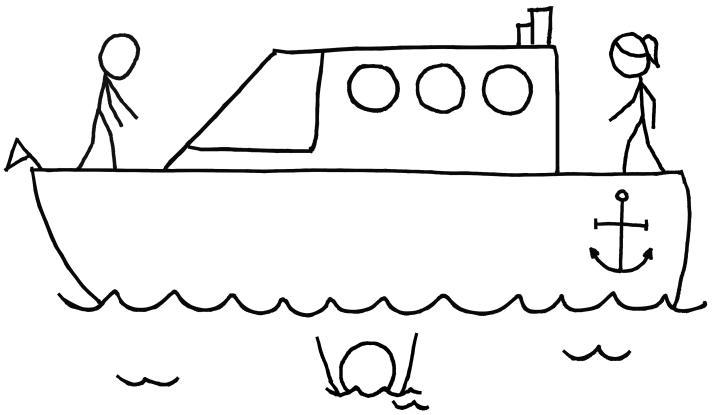


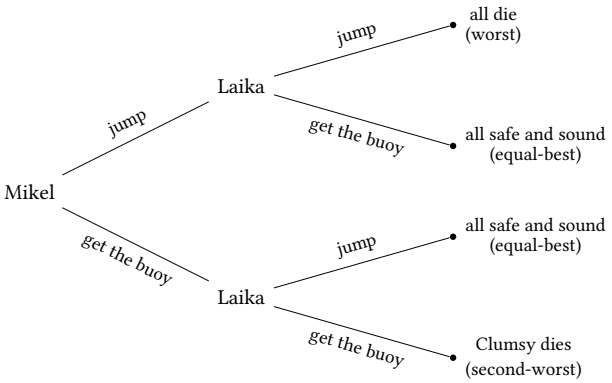
Figure 8.2: The situation from Seaman Clumsy.

Seaman Clumsy is not only a symmetric Coordination Case, but also an interesting one, because the only assessments we get, according to the traditional analysis we reconstructed, are conditional assessments, i.e.,

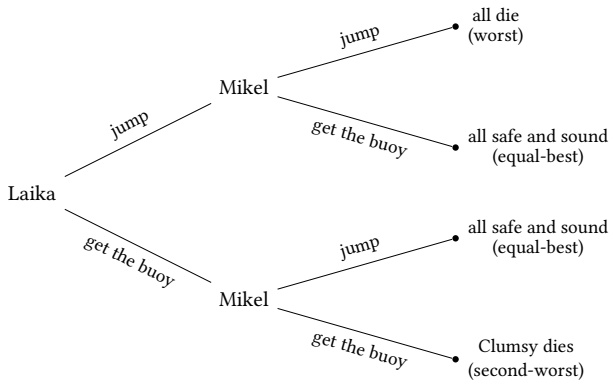
(48) If Mikel (Laika) jumps, it is right for Laika (Mikel) to get the buoy.

(49) If Mikel (Laika) gets the buoy, it is right for Laika (Mikel) to jump.

However, if we adopt Approach, we find that both Mikel and Laika are in valuatively identical decision situations, namely



and



Alternatively, we can can make use of our new representational vehicle, the GEF. Figure 8.3 illustrates the GEF for Seaman Clumsy. At a glance, we can see that whichever agent acts first places the second-acting agent in a *valuatively identical* decision situation. For instance, if Laika, as the first-acting agent, jumps, Mikel is left with a choice between bringing about one of the equally best outcomes (by fetching the buoy) or causing the worst outcome (by also jumping). The same applies if their roles are reversed. Similarly, if Laika first fetches the buoy, Mikel then faces the decision between bringing about one of the equally best outcomes (by jumping) or causing the second-worst outcome (by also fetching the buoy). Again, this holds true in reverse. Consequently, the following is guaranteed: regardless of which option (or options) MOAC deems right for the first-acting agent, it should, in a meaningful sense, prescribe the same type of option(s) for both agents. The precise nature of this 'should' is clarified in the following discussion.

As an observation, note first that when discussing an objective consequentialist assessment of a case like Seaman Clumsy, it does *not* matter how the outcomes are valuated exactly. Once we assume the moral assessment is objective consequentialist—i.e., it depends solely on the moral qualities of the outcomes, not, for instance, on which agent happens to be in which situation—the respective options of the second-acting agent in Seaman Clumsy must be valuated or ranked equally. The same then holds for the valuation of the decision situation in which this agent finds themselves, as this is, as discussed in the last chapter, a function of the values of the final outcomes. For example, suppose the value assigned to the upper action situation (where a jump would lead to the death of all involved) is V_1, while the value assigned to the bottom action situation (where a jump would ensure the survival of all) is V_2. Then we are left with the following situations to consider:

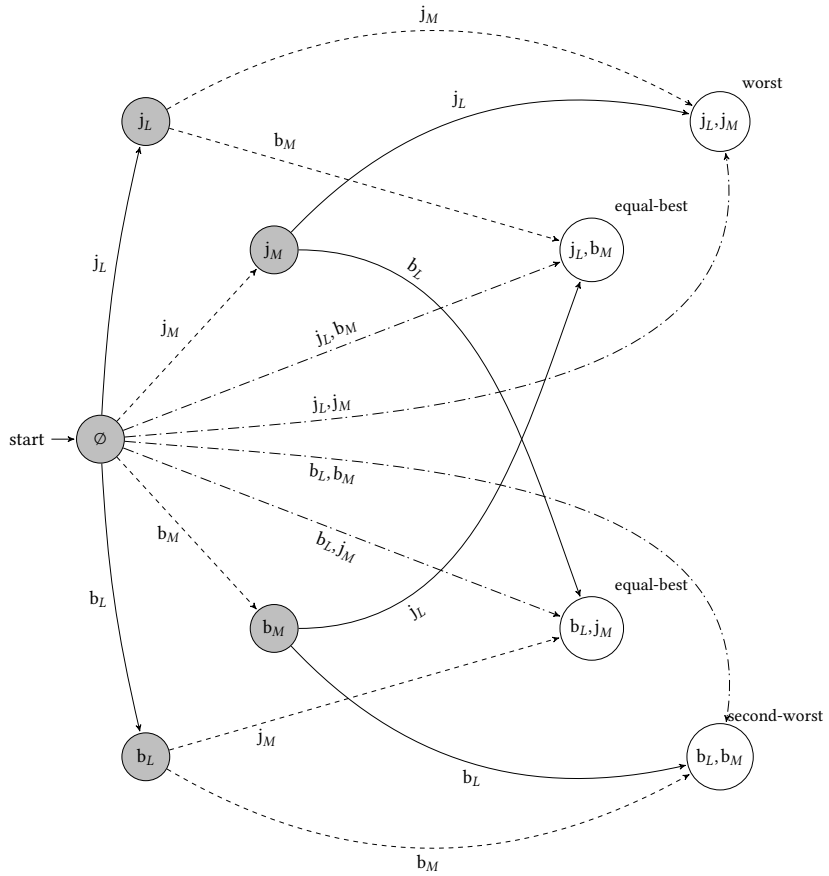
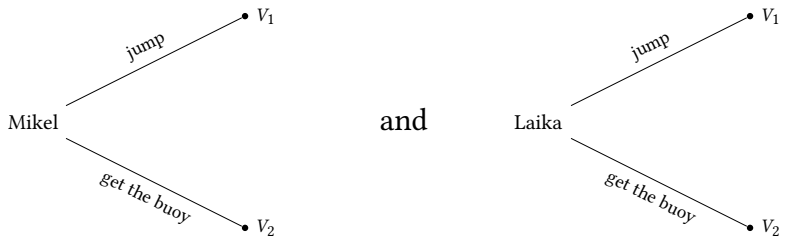


Figure 8.3: Collective Extensive Form of Seaman Clumsy (where j stands for jumping, b for getting the buoy, *M* for Mikel and *L* for Laika).



It should be obvious that no objective act-consequentialist theories can come to different assessments in these two cases. This can also be explicitly shown. Recall:

Definition 2.3 (Objective Consequentialist Theory (formal)) *T is an* objective consequentialist theory *if and only if it embraces an axiological subtheory T_{Ax} with a valuation function* $\mathsf{Val}: \mathcal{W} \to \mathcal{V}$ *and an objective consequentialist criterion of rightness T_{CoR} such that, for all decision situations $D \in \mathfrak{I}$ and for all $\phi \in \Phi_D$: $D, C \vDash_T \mathrm{R}\phi$ if and only if $T_{\mathrm{CoR}}(\phi)$.*

A criterion of rightness T_{CoR} is objective consequentialist *if and only if, for all* $D \in \mathbb{I}$ *with* $\mathsf{D} := \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O}\rangle$ *(with* D*'s actual context* C*)* T_{CoR} *corresponds to a predicate* $\chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}$ *such that for all* $\phi \in \Phi$*:*

$$D,C \vDash_T \mathrm{R}\phi \quad \textit{if and only if} \quad \chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}(\mathsf{Val}(\mathsf{Out}_C(\phi))).$$

Indeed, this definition *implies* that any objective consequentialist theory has another property that I will call *normative supervenience.*[148] Roughly speaking, for a theory to have this property means that, in valuatively identical decision situations, the theory necessarily assigns the same moral status to corresponding options. We can explicate the involved terms.[149] We first define:

Definition 8.1 (Valuative Embedding) *Let* $D, D' \in \mathcal{D}$ *be two individual decision situations with*

$$\mathsf{D} = \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O}\rangle$$

and

$$\mathsf{D}' = \left\langle A', \Phi', \mathsf{Out}'_{C'} : \Phi' \to \mathcal{O}\right\rangle$$

and let $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$ *be a valuation function.* D *can be* valuatively embedded *in* D' *(relative to* Val*) if and only if there is a injection* $f : \Phi \to \Phi'$ *such that for all* $\phi \in \Phi$*:*

$$\mathsf{Val}(\mathsf{Out}_C(\phi)) = \mathsf{Val}(\mathsf{Out}'_{C'}(f(\phi))).$$

Next, we define:

Definition 8.2 (Valuative Identity) *Let* $D, D' \in \mathcal{D}$ *be two individual decision situations with*

$$\mathsf{D} = \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O}\rangle$$

and

$$\mathsf{D}' = \left\langle A', \Phi', \mathsf{Out}'_{C'} : \Phi' \to \mathcal{O}\right\rangle$$

and let $\mathsf{Val} : \mathcal{W} \to \mathcal{V}$ *be a valuation function.* D *is* valuatively identical *to* D' *(relative to* Val*) if and only if* D *can be valuatively embedded in* D' *and vice versa.*

[148] The thoughts presented here are my own. However, I later found out that Krister Bykvist had the same insight more than twenty years ago, cf. Bykvist 2002, 2003. The similarities go so far that he also used the term "Normative Supervenience" (this is even how I found his article) in this context, even though the properties presented here are rather what he calls "Consequentialist Supervenience".

[149] This is the moment promised in Subsubsection 3.5.2.1: we now have to precisely define the notion of structural equivalence captured there only intuitively.

This is to say that for every option in the one decision situation, there is an option with an equally good outcome in the other decision situation. Thus, for valuatively identical decision situations, we have an injection from the one decision situation's option space into the other's decision situation's option space and vice versa, and so we know that there is a bijection between these option spaces.[150] We call such a bijection a *correspondence relation.*

Based on these notions, we can define

Property 8.1 (Normative Supervenience) *Let T be a moral theory and D a decision situation. The moral assessments of T* normatively supervene on *the valuative profile of D if and only if, necessarily, for every decision situation D' that is valuatively identical to D (relative to a valuation function $\mathsf{Val}: \mathcal{W} \to \mathcal{V}$ and a correspondence mapping $f : \Phi_D \to \Phi_{D'}$), the following holds:*

For every $\phi \in \Phi_D$ and $\phi' \in \Phi_{D'}$, if $f(\phi) = \phi'$, then ϕ and ϕ' have the same moral status according to T in their respective decision situations.

It is indeed easy to see that every objective consequentialist moral theory (in accordance with Definition 2.3) has Normative Supervenience.[151]

This brings us back to Seaman Clumsy. Mikel and Laika's decision situations are valuatively identical since we can map Mikel's option to jump to Laika's option to jump, and his option to get the buoy to her option to get the

[150]Note that we assume the option spaces are finite, ensuring that mutual embeddings imply a bijection. This follows from the fact that injective mappings between finite sets of the same cardinality are bijections. Additionally, the valuative conditions are preserved in both directions, ensuring consistency in the mappings between the two option spaces. For infinite option spaces, however, further justification or structural assumptions would be required to guarantee bijectivity.

[151]The proof is easy, but a bit lengthy: Let T be an objective consequentialist theory and let $\mathsf{D} = \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O}\rangle$ and $\mathsf{D}' = \left\langle A', \Phi', \mathsf{Out}'_{C'} : \Phi' \to \mathcal{O}\right\rangle$ (where C is the actual context of D and C' the actual context of D') be the representations of two valuatively identical decision situations (relative to valuation function $\mathsf{Val}: \mathcal{W} \to \mathcal{V}$ and correspondence mapping $f : \Phi_D \to \Phi_{D'}$). To prove the claim, we need to show that for every $\phi \in \Phi_D$ and $\phi' \in \Phi_{D'}$: If $f(\phi) = \phi'$, then

$$D, C \vDash_T R\phi \quad \text{if and only if} \quad D', C' \vDash_T R\phi'.$$

Since T falls under Definition 2.3, there is a predicate $\chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}$ such that for all $\phi \in \Phi$:

$$D, C \vDash_T R\phi \quad \text{if and only if} \quad \chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}(\mathsf{Val}(\mathsf{Out}_C(\phi)))$$

and there is a predicate $\chi_{T,\mathsf{Val}(\mathcal{O}_{D',C'})}$ such that for all $\phi' \in \Phi'$:

$$D', C' \vDash_T R\phi' \quad \text{if and only if} \quad \chi_{T,\mathsf{Val}(\mathcal{O}_{D',C'})}(\mathsf{Val}(\mathsf{Out}'_{C'}(\phi'))).$$

Thus, it suffices to show that, if $f(\phi) = \phi'$, then

$$\chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}(\mathsf{Val}(\mathsf{Out}_C(\phi))) \quad \text{if and only if} \quad \chi_{T,\mathsf{Val}(\mathcal{O}_{D',C'})}(\mathsf{Val}(\mathsf{Out}'_{C'}(\phi')))$$

buoy. Thus, since MOAC *is* an objective consequentialist theory—of which we have long since assured ourselves in Chapter 2—whatever moral status MOAC assigns to one of these options it has to assign to the other. Given the standard assumptions with respect to the values V_1 and V_2 (most importantly, the assumption that they are comparable and, thus, that we can rank them) we have four obvious possible assessments to consider:

- Both options—to jump and to get the buoy—are meant to be right according to an improved version MOAC. Given these assessments, it cannot be guaranteed that Mikel and Laika, when acting in accordance with MOAC, bring about the best outcome that they could together bring about. It could well be that the one jumps and that the other gets the buoy; but they might also both get the buoy or both jump. The latter combinations are suboptimal and one even results in the *worst* possible outcome, namely the death of all three.

- Alternatively, such a version of MOAC might assess only the option of jumping as being right. Then Mikel and Laika, if they acted according to MOAC, would not only *miss* the optimal results; they would even with certainty produce the *worst* possible result.

- Finally, such a version of MOAC could asses only the option to get the buoy as right. Then Mikel and Laika, if they acted accordingly, would miss the optimal result, but they would at least be guaranteed to produce the second-best result.

Now, since D and D' are valuatively identical decision situations (relative to valuation function $\mathsf{Val}: \mathcal{W} \to \mathcal{V}$ and correspondence mapping $f : \Phi_D \to \Phi_{D'}$), we know several things that help us to establish this equivalence. First, we know that for all $\phi \in \Phi$: $\mathsf{Val}(\mathsf{Out}_C(\phi)) = \mathsf{Val}(\mathsf{Out}'_{C'}(f(\phi)))$. Thus, since, by assumption, $f(\phi) = \phi'$, we know that $\mathsf{Val}(\mathsf{Out}_C(\phi)) = \mathsf{Val}(\mathsf{Out}'_{C'}(\phi'))$. Second, we know that $\mathsf{Val}(\mathcal{O}_{D,C}) = \mathsf{Val}(\mathcal{O}_{D',C'})$ because we defined

$$\mathsf{Val}(\mathcal{O}_{D,C}) := \{\, \mathsf{Val}(O) \in \mathcal{V} \mid O \in \mathcal{O}_{D,C} \,\},$$

which we can rewrite (given that, by definition, $\mathcal{O}_C = \{\, \mathsf{Out}_C(\phi) \mid \phi \in \phi \,\}$ for an individual decision function D with option space Φ and outcome function Out_C) as

$$\mathsf{Val}(\mathcal{O}_{D,C}) = \{\, \mathsf{Val}(\mathsf{Out}_C(\phi)) \in \mathcal{V} \mid \phi \in \Phi \,\},$$

which, in light of the above first established identities and the existence of bijection f, we can rewrite $\mathsf{Val}(\mathcal{O}_{D,C})$ as $\{\, \mathsf{Val}(\mathsf{Out}'_{C'}(\phi')) \in \mathcal{V} \mid \phi' \in \Phi' \,\}$. Thus, $\mathsf{Val}(\mathcal{O}_{D,C}) = \mathsf{Val}(\mathcal{O}_{D',C'})$. Therefore, by substituting both the parameter and the argument of χ, it holds indeed that

$$\chi_{T,\mathsf{Val}(\mathcal{O}_{D,C})}(\mathsf{Val}(\mathsf{Out}_C(\phi))) \quad \text{if and only if} \quad \chi_{T,\mathsf{Val}(\mathcal{O}_{D',C'})}(\mathsf{Val}(\mathsf{Out}'_{C'}(\phi'))).$$

In other words: If T is a objective consequentialist theory, then it adheres to Normative Supervenience.

- In light of the lessons of the last chapter, there is a fourth possibility. Such a version of MOAC could command both to jump with a certain probability (greater than 0 and less than 1) and to fetch the buoy with the opposite probability. But even then, achieving the best result is just *possible* but not guaranteed. After all, the probability for the respective corresponding options must be equal (i.e., MOAC would need to assign the same probability distributions over the options of both agents, relative to some correspondence mapping); that's what Normative Supervenience requires. Therefore, even adopting such a *mixed* strategy cannot lead to an optimum with certainty.

My point here is this: even *without* having thought about how to assess or rank the individual decision situations that Mikel and Laika can produce in Seaman Clumsy in concrete terms, we already know that *if* they are 'reasonably' ranked (that is, in line with other traditional consequentialist assumptions and Normative Supervenience), they are *guaranteed* to be assessed by MOCOR in a way that *allows* for suboptimal outcomes.

In other words, adopting the APPROACH allows us to resolve the REAL CHALLENGE, but it still falls short of fully satisfying MH. However, this does *not* imply that the APPROACH itself is flawed simply because it reintroduces the CHALLENGE in cases like Seaman Clumsy (even though it can handle Troublemakers effectively, as we will discuss later). Instead, one thing should be clear: adhering to PMH *in the form of MH* (or in any of the similar formulations we have encountered thus far) for cases like Seaman Clumsy amounts to *abandoning* objective consequentialism. The reason is clear: in such cases, every objective-consequentialist theory inevitably conflicts with MH due to its logical commitment to Normative Supervenience. This is the central takeaway of this section.

What we need, therefore, to effectively address the task of valuating the newly discovered consequences of actions in interesting collective decision situations is a *reasonable and widely acceptable explication* of the general idea behind PMH. This explication should then serve as the benchmark for closing the gaps.

8.1.2 Upshot: Reasonable Moral Harmony

It should be evident by now that a straightforward formalization of the collective reading of Congruence—such as a naïve version of PMH like MH—inevitably conflicts with other, indisputable principles in certain cases. What is needed, therefore, is a more nuanced formulation of PMH that accommodates additional requirements for moral theories. Such a formulation must avoid violating other well-founded or even theory-entailed constraints and, above all, must not demand the impossible from objective consequentialist

theories. This need was already anticipated in the discussions of Broad's Property (cf. Subsection 3.5.1) and Pinkert's On-the-hook (cf. Subsubsection 3.5.2.2, particularly page 85). Consequently, we are prepared to develop a less subjective and more refined version of these properties.

Principle 8.1 (Collectively Maximizing– Reasonable) *If all agents act (consistently) rightly, then they are guaranteed to produce the morally best outcome they can* reasonably *be expected to collectively bring about.*

This formulation naturally raises at least two critical questions. First, what does it mean, in light of the Objective View, to say that a particular outcome *is to be expected*? Shouldn't everything already be set and determined? The answer, of course, is no. Methodological Indeterminism inevitably introduces uncertainty whenever multiple agents' actions jointly determine which specific outcome will occur. This principle, as we have seen in the last two chapters, ensures that not everything is predetermined. However, the details of this uncertainty remain to be fully worked out.

Second, we must ask what it means to expect an outcome *reasonably*. This part of the formulation clearly aims to hedge against exceptions, such as Seaman Clumsy or, more generally, cases with non-unique global optima, which challenge the applicability of PMH. The pressing question, then, is how to understand and articulate this hedge in a more precise way.

In the following section, I will address both of these questions. This will lead us into the domain of formal decision theory, a field offering a rich repertoire of concepts, notions, and methods. These tools will not only help us clarify the formulation on both counts but will also illuminate a path for addressing the gaps posed by the REAL CHALLENGE—without allowing the CHALLENGE to reemerge.

8.2 Amendments for Reasonable Pathfinding

The task at hand is to determine precisely what it means that the right action of all agents guarantees the morally best outcome that can reasonably be expected.

We proceed in two steps. First, we develop a precise understanding of what it means to act (consistently) right in this collective sense—that is, what it means for everyone to act rightly on all occasions. The basic idea is as follows: Presuming the APPROACH, we consider the notion of a function π that systematically selects one of the options available to an agent within a given individual decision situation $D \in \mathbb{I}$. This function determines which actions are performed based on the decision situation. Such functions, known as *policies*, are well established in formal decision theory, particularly within the

frameworks of Markov decision processes (MDPs)—formal models for sequential decision-making under uncertainty—and reinforcement learning. Policies are tools that model systematic (or at least consistent) decision-making. While traditionally explored in the domain of instrumental rationality, policies offer a formal yet flexible framework that can be adapted for modeling moral theorizing. The aim of the following section is to leverage the concept of policies to address the central task of this project and to expand their applicability to (moral) consequentialist frameworks more broadly.

Second, we need to establish a criterion for what makes a policy *reasonably defensible* in light of an informed version of PMH—and, subsequently, what makes one reasonably defensible policy preferable to another from the perspective of MOAC. This framework provides a method for valuating policies, which we will use in the second part of this chapter to assess various collective amendments. As we will see, MOAC, in combination with any reasonable and defensible amendment, implies a reasonably defensible policy.

8.2.1 On Policies and Their Evaluation

In formal decision theory, particularly within the framework of MDPs and prominently in reinforcement learning, the term "policy" carries a specific technical meaning. While the term might initially evoke a general notion of guidelines or strategies, its definition in this context is far more precise. A policy, typically denoted by the symbol "π", is a function that determines an agent's actions in every possible state or decision situation.

Policies can be deterministic or indeterministic. Deterministic policies are functions $\pi : \Phi \to \mathcal{W}$ that select a single option in a given decision situation. Indeterministic policies, on the other hand, are probability distributions $\pi : \Phi \to [0,1]$, specifying the likelihood of an agent performing a particular action. Since indeterministic policies generalize deterministic ones—and given that the probabilistic amendment (Mixed Strategies) is still under consideration and can only be modeled with indeterministic policies—I will focus on probabilistic policies in the following discussion.

The goal in many MDPs and reinforcement learning scenarios is to identify an *optimal* policy, often denoted as π^*, which maximizes the agent's *expected* reward over time. Analogously, a moral consequentialist policy can be thought of as a guideline for consistent moral right-doing, directing an agent's actions in each relevant moral scenario to maximize the expected moral value over time. The connection to the concept of *reasonable* Collectively Maximizing is then quite apparent: *defensible* moral theories correspond to *reasonable* policies.

What we seek in this chapter is, ultimately, a policy that is *as good as possible* in the sense of MOAC—one that ensures the morally best possible

outcomes if all agents adhere to it strictly. To systematically explore this, we begin by formalizing the concept of a policy within the framework of a collective decision situations (or, more precisely, given the scope of this project for maximal and order-invariant Coordination Cases).

We define policies over generalized extensive forms (GEFs) of such decision situations. Recall from Definition 7.1 that the GEF of a decision situation D is a tuple $\mathcal{G}(D) = \langle \mathcal{S}, \mathcal{E}, S_\emptyset \rangle$, where:

- $\mathcal{S}$ is the set of states, including intermediate and final states.
- $\mathcal{E}$ is the set of transitions, representing sequences of actions by which D is resolved.
- $S_\emptyset$ is the initial state.

Additionally, we defined the set of intermediate states ($\mathcal{S}^{\text{inter}} \subset \mathcal{S}$) as the subset of states corresponding to partial combinations of actions. For any such intermediate state S_Υ, the decision situation $D_{\downarrow\Upsilon}$ represents the residual decision situation after Υ is performed.

To capture the full range of possible actions across all stages of D, we define

$$\Phi^D := \bigcup_{S_\Upsilon \in \mathcal{S}^{\text{inter}}} \Phi_{D_{\downarrow\Upsilon}},$$

denoting the options available at S_Υ. This set represents the entire option space for D as it unfolds. Φ^D forms the foundation for defining policies. Φ^D helps us to concisely define policies.

We now introduce the concept of a *policy* as a mapping that assigns probabilities to actions in Φ^D while satisfying the standard axioms of probability at every intermediate state:[152]

Definition 8.3 (Policies for GEFs) *Let $\mathcal{G}(D) = \langle \mathcal{S}, \mathcal{E}, S_\emptyset \rangle$ be the generalized extensive form of decision situation D. A* policy for D *is a function $\pi : \Phi^D \to [0,1]$ such that, for every intermediate state $S_\Upsilon \in \mathcal{S}^{inter}$, $\pi(\phi)$ specifies the probability of selecting $\phi \in \Phi_{D_{\downarrow\Upsilon}}$ at S_Υ.*

Since we can transition between the GEF and its corresponding decision situation D, we may equivalently refer to a policy for a decision situation D—as long as D can be expressed as a GEF (recall that all decision situations

[152] This means that a policy $\pi : \Phi^D \to [0,1]$ fulfills the standard axioms for probability distributions over finite (or, more generally, discrete) sample spaces at every intermediate state $S_\Upsilon \in \mathcal{S}$, i.e.,

$$\forall \phi \in \Phi_{D_{\downarrow\Upsilon}} : 0 \leq \pi(\phi) \leq 1 \quad \text{and} \quad \sum_{\phi \in \Phi_{D_{\downarrow\Upsilon}}} \pi(\phi) = 1.$$

relevant to the present project satisfy this condition).[153] Most importantly, by definition, a policy π for a decision situation D is also a policy for every decision situation D' that is an intermediate outcome of D and to distinguish the specific policy for D' from π I will refer to that induced policy $\pi_{D'}$. Specifically, a policy for a collective decision situation D is also a policy for every decision situation in the decomposition $\mathbb{I}_D := \{ \mathsf{D}_A \mid A \in \mathcal{A} \}$ of D.

Next, we define what it means for a policy to apply to an entire set of decision situations, whether individual or collective:

Definition 8.4 (Domains of Policies) *Let $\mathfrak{D}_\pi$ be a set of decision situations (individual or collective) that can be expressed in terms of generalized extensive forms. $\mathfrak{D}_\pi$ is the* domain of a policy *π if and only if π is a policy for every $D \in \mathfrak{D}_\pi$.*

If a set $\mathfrak{D}_\pi$ is the domain of a policy π, we also say that π is a policy for the domain $\mathfrak{D}_\pi$.

Before addressing the relationship between policies and moral theories, how policies can help evaluate moral theories, and what is still required to derive policies effectively from MOAC theories (including potential amendments), it is crucial to clarify two foundational aspects: first, how to evaluate policies, and second, how to rank policies.

While the intuitive notion of evaluating a policy π relative to a collective decision situation D (or a set of such situations) is straightforward, formalizing this intuition poses challenges. Conceptually, to determine the expected value of π, we need to calculate the probability of each possible combination of actions $\Upsilon \in \Psi_D$ according to π—denoted as $\pi(\Upsilon)$. This probability is then multiplied by the value assigned to that combination, and the results are summed. This approach gives rise to the following lifting of valuation functions to the policy level:

Definition 8.5 ((Expected) Value of a Policy) *Let D be a collective decision situation with $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi \to \mathcal{O} \rangle$ and actual context C, let π be a policy for D, and let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be a valuation function. The (expected)* value of π relative to D and Val *is defined as:*

$$\mathsf{Val}_C(\pi) := \sum_{\Upsilon \in \Psi} \pi(\Upsilon)\mathsf{Val}_C(\Upsilon).$$

Calculating $\pi(\Upsilon)$, however, becomes complex, especially when the order of actions is taken into account. At a minimum, we need a framework for

[153] Although this book restricts itself to maximal and order-invariant Coordination Cases, the concept of GEF can be generalized to non-maximal and non-order-invariant Coordination Cases. Extending the formal framework in this way is left, again, as an exercise for interested readers.

determining the probability of which agent acts next. One simple assumption could be that each agent is equally likely to act.[154] This scenario is modeled as a uniform probability distribution over the decomposition of a collective decision situation D, i.e., $\mathfrak{I}_D := \{ \mathsf{D}_A \mid A \in \mathcal{A} \}$. Under this assumption, the probability of an agent A acting next is:

$$\Pr_{\mathfrak{I}_D}(\mathsf{D}_A) = \frac{1}{|\mathcal{A}|}.$$

This uniform distribution is a straightforward solution, but it does not account for cases of simultaneous actions. Incorporating such cases would add considerable complexity and require more substantial modeling assumptions.[155]

For the purposes of this project, however, we can simplify significantly. As the subsequent sections will show, the different amendments (in combination with MOCOR) provide relatively clear guidance on the probabilities with which agents in the simplest collective decision situations are expected to perform specific actions. This allows us to approximate $\pi(\Upsilon)$ without undue complexity. Nonetheless, it remains essential to pay close attention to the edge case of simultaneous actions, as these could (and, as I am going to argue, will) have significant implications for the evaluation process.

The definition of ranking policies from the consequentialist point of view is then very straightforward:

Definition 8.6 (Consequentialist Policy Ranking) *Let D be a decision situation (either individual or collective) with actual context C, let π and π' be two policies for D, and let* $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ *be some valuation function. The policy π is* preferred (from the consequentialist point of view) to π' relative to D and Val *(in symbols, $\pi >_{C,D} \pi'$) if and only if*

$$\mathsf{Val}_{C,D}(\pi) > \mathsf{Val}_{C,D}(\pi').$$

Based on these notions, we can finally turn to the questions of how to derive policies from moral theories and how this enables camp MOAC decide on the "right" amendments.

[154] Alternatively, one could introduce a bias in this distribution based on some kind of meta-amendment. While I cannot think of a compelling reason to exploit this additional degree of freedom and thus do not pursue it here, it remains a potential avenue for future exploration by camp MOAC.

[155] One approach could involve assigning a timer to each agent with an associated probability distribution, determining the likelihood of an agent acting at any given moment. Similar concepts have been explored in computer science, particularly in the context of Concurrent Markov Decision Processes (CoMDPs) and even with continuous time models in Continuous-time Markov Decision Processes (CTMDPs) (cf. Guo et al. 2009).

8.2.2 Evaluating Amendments

Policies do not arise out of thin air. In the context of moral theorizing, it is natural to view them as expressions of moral theories. Indeed,—at least once we introduce one final theoretical component—certain policies can be seen as direct and immediate expressions of MOAC theories. We establish this connection in two steps.

First, we define what it means for a policy to be *consistent with the criterion of rightness of a theory* T (or, in short, *consistent with* T). A policy π is consistent with T if and only if π is a policy for T's domain $\mathfrak{I}_T$, and for all individual decision situations $D \in \mathfrak{I}_T$ (with actual context C) and for all $\phi \in \Phi_D$:

$$\pi(\phi) > 0 \quad \text{if and only if} \quad T, D, C \vDash R(\phi).$$

In other words, a policy consistent with T assigns non-zero probability only to those options deemed *right* by T's criterion of rightness, and it excludes all others. Agents who act in accordance with a policy consistent with T can never violate T; conversely, they will *necessarily always satisfy* T.

For classical, deterministic moral theories—and under the assumption that exactly one option is right in D given C according to T—the above condition simplifies to:

$$\pi(\phi) = 1 \quad \text{if and only if} \quad T, D, C \vDash R(\phi).$$

But what if there is more than one right action? It is generally assumed that, in such situations, an agent can *arbitrarily* choose from the set of right actions (cf. Section 2.2).[156]

This brings us, second, to yet another possible probabilistic element, one that is seldom discussed in normative ethics. As announced, this probabilistic element is not to be accompanied by heavyweight, metaphysical determinations but to be of an analytical nature. It comes from a rather trivial enrichment of our concept of a full-fledged normative theory: a truly complete moral theory should offer not only a criterion of rightness (and, in the case of consequentialist theories, also an axiological module and a relevance stance, cf. Section 2.3) but also a *selection rule* that determines, for a set of options assessed as right according to its criterion of rightness, how, ultimately, the option actually performed is permissible to select.

[156]This is where instrumental rationality and morality might intersect: when an agent faces no further moral constraints within the set of right actions, they could reasonably choose based on their personal interests or inclinations—or alternatively, they might decide to set these aside, potentially acting in a supererogatory manner (cf. Wessels 2002). However, in the cases considered here, such inclinations, interests, or dispositions are not explicitly stated and therefore cannot resolve the decision. Consequently, I assume that no preference ordering exists beyond what is already accounted for in the moral assessment of the outcomes.

This brings us, secondly, to another potential probabilistic element, one that is rarely discussed in normative ethics. As previously noted, this element does not rely on heavyweight metaphysical commitments but instead serves an analytical purpose. It arises from a straightforward yet often overlooked enrichment of the concept of a fully developed normative theory. A *truly* comprehensive moral theory should not only include a criterion of rightness (and, for consequentialist theories, an axiological module and a relevance stance, cf. Section 2.3) but also a *selection rule*. This rule determines, within the set of options deemed right according to the criterion of rightness, how the specific option ultimately chosen may permissibly be selected.

To illustrate such a moral version of Buridan's ass in a collective setting, consider again the following normal form of the modified Whiff and Poof example:

		Poof	
		not-push	push
Whiff	not-push	10	0
	push	0	10

Relative to all the amendments discussed so far that are still in play, the two agents here face equally good options. Since we are searching for theories that adhere to No Moral Dilemmas (cf. Subsection 4.3.1) and Normative Supervenience, both actions should better be deemed right by our preferred MOAC theory (for which we are still searching). Assuming that any preferences of the agents are already represented in the valuation of outcomes, what can we reasonably expect from the actions of agents who always act rightly in the sense of MOAC?

I think that, in such cases of moral indifference among equally right options and under the assumption of complete indifference of the agent,[157] it is reasonable to propose that MOAC employs a tie-breaker in terms of *fair randomization*. Specifically, MOAC theories could adopt a simple probabilistic selection rule that assigns equal probability to each morally right action.[158] Formally, for any decision situation $D \in \mathfrak{I}_T$ (with actual context C) and an arbitrary MOAC theory T, we define the probability distribution that serves as the corresponding selection rule as:

$$\Pr^T_{D,C} : \Phi_D \to [0,1], \quad \Pr^T_D(\phi) = \begin{cases} \frac{1}{|T(D,C)|}, & \text{if } \phi \in T(D,C), \\ 0, & \text{otherwise.} \end{cases}$$

[157] Cf. footnote 156.

[158] Like with the question of the likelihood of which agent acts next introduced in the last subsubsection, I again see no reason for anything but uniform distributions, but see footnote 154.

This now allows us to define:

Definition 8.7 (Theory-Induced Policy) *Let T be a normative theory. A policy π_T is called a* policy induced by *T (also called a T-*induced policy*) if and only if π_T is consistent with T and π_T mirrors T's selection rule.*

For a MOAC theory T that does not allow for mixed strategies, we then have that π_T is a T-induced policy if and only if for all $D \in \mathbb{I}_T$ (with actual context C) and all $\phi \in \Phi_D$,

$$\pi_T(\phi) > 0 \quad \text{if and only if} \quad T, D, C \vDash R(\phi),$$

and

$$\pi_T(\phi) = \Pr^T_{D,C}(\phi).$$

This means that we can freely switch between any MOAC theory T and its induced policy, enabling us to finally rank amendments.

Definition 8.8 (Consequentialist Amendment Ranking) *Let T be a moral theory with normative gaps, and let Δ_1 and Δ_2 be two amendments that fill these gaps. Let T_1 and T_2 denote the moral theories resulting from T adopting Δ_1 and Δ_2, respectively. Finally, let D be a decision situation (either individual or collective) with actual context C.*

Δ_1 is to be preferred (from the consequentialist point of view) *to Δ_2 relative to D (in symbols, $\Delta_1 >_{C,D} \Delta_2$) if and only if*

$$\pi_{T_1} >_{C,D} \pi_{T_2}.$$

Based on these definitions, we can now rank different amendments relative to individual and collective decision situations. Thus, it only remains to execute this program for the amendments still in play and with respect to the decision situations most relevant to this work.

8.3 The Final Evaluation

We now have everything on the workbench to decide on the remaining amendments. The candidate amendments (see again Table 8.1) are in place; we have a framework for assessing them in terms of their "performance", specifically by valuating the policies they imply (in interaction with MOCOR and the selection rule established earlier); and we have identified two interesting and challenging types of Coordination Cases to serve as "testbeds" (Troublemakers and cases with the structure of Seaman Clumsy).

It is important to acknowledge the limitations of this assessment. The results cannot claim universal generality: there may be amendments beyond

Amendment	Remarks	Still in the Game?
Summation	promising	in
Maximization	promising	in
Expected Utility	Although in principle it is not impossible to do something with it, the assumption of objective probabilities brings too great a burden with it. It might be a bigger project to fill this approach with life. Put aside for the time being.	out
Dominance	Allows moral dilemmas and otherwise offers nothing that we could not already derive with the Sure-Thing Principle.	out
Non-Dominated	Too permissive, especially in combination with PMH-driven, optima-demanding intuitions (and this is where we are coming from).	out
MaxiMin	Too permissive in combination with PMH-driven, optima-demanding intuitions.	out
MaxiMax	Extensionally equivalent to Maximization but comes with modification of MOCOR.	out
Mixed Strategies	promising	in

Table 8.1: The introduced amendments, which of them are still in the game, and which are no longer (and, in brief, why the amendments that were thrown out.

the scope of this investigation, as well as other significant (types of) Coordination Cases that merit consideration. Nevertheless, the CHALLENGE will be meaningfully addressed by our focus on Troublemakers, and by examining Seaman Clumsy–like cases, we account for the implications of PMH and Normative Supervenience.

What remains is to conduct the evaluation and render a final judgment. These represent the concluding steps of this project.

8.3.1 Defining a Testbed

The proposed methodology for assessing possible amendments for MOAC theories is based on assessing the policy resulting from MOCOR together with the respective amendment relative to *evaluative test cases*, i.e., relative to a set of decision situations. Since the central goal of this work is to master the CHALLENGE for MOAC, classical Troublemakers (with the structure of Whiff and Poof or Two Factories) should definitely belong to the set of evaluation test cases. Furthermore, cases like Seaman Clumsy should also be part of it, because they can serve as touchstones with respect to performance on

cases with more than one optimum and especially with respect to Normative Supervenience.

Thus, we will consider two types of Coordination Cases,[159] defined through their evaluative profiles. The first type, denoted by trouble (with decomposition $\mathfrak{I}_{\text{trouble}}$), is given by the (structural) normal form of Two Factories–like cases (with $v_{12} = v_{21} < v_{11} < v_{22}$), recall:

trouble		Y	
		ϕ_Y	$\neg\phi_Y$
X	ϕ_X	$-\ (v_{11})$	$---\ (v_{12})$
	$\neg\phi_X$	$---\ (v_{21})$	$+++\ (v_{22})$

The second type, denoted by sailor (with decomposition $\mathfrak{I}_{\text{sailor}}$), is defined by the (structural) normal form of Seaman Clumsy–like cases (with $v_{11} < v_{21} < v_{12} = v_{21}$), recall:

sailor		Y	
		ϕ_Y	$\neg\phi_Y$
X	ϕ_X	$---\ (v_{11})$	$+++\ (v_{12})$
	$\neg\phi_X$	$+++\ (v_{21})$	$-\ (v_{22})$

Annotated GEFs of trouble and sailor can be found in Figure 8.4 on the next page.

A wide range of criteria for evaluating good solutions has been developed throughout this book and now come together in this section (cf. especially Chapter 4). First, the goal is to identify a truly objective-consequentialist theory in the sense of Definition 2.3. More specifically, we seek a MOAC theory that, crucially, must be consistent.

Second, a reasonably acceptable version of PMH (as defined in Principle 8.1 in Subsection 8.1.2) serves as a general guideline for selecting suitable amendments, as explained in detail in Section 8.2, where policies are used as the key evaluative framework.

Third, the aim is to satisfy No Moral Dilemmas and, as far as possible, achieve Deontic Completeness, as emphasized in Section 4.3. Addressing the deontic gaps diagnosed in Section 6.5 as the root cause of the REAL CHALLENGE is critical. These gaps must be closed in a manner that allows agents to act rightly, avoiding the introduction of moral dilemmas, and aligning with Resolvability.

[159] As already mentioned (cf. Footnote 137, see also Figure 7.9), Sequential Cases should automatically be treated in parallel because, against the background of the APPROACH, each Coordination Case is ultimately a superposition of multiple Sequential Cases.

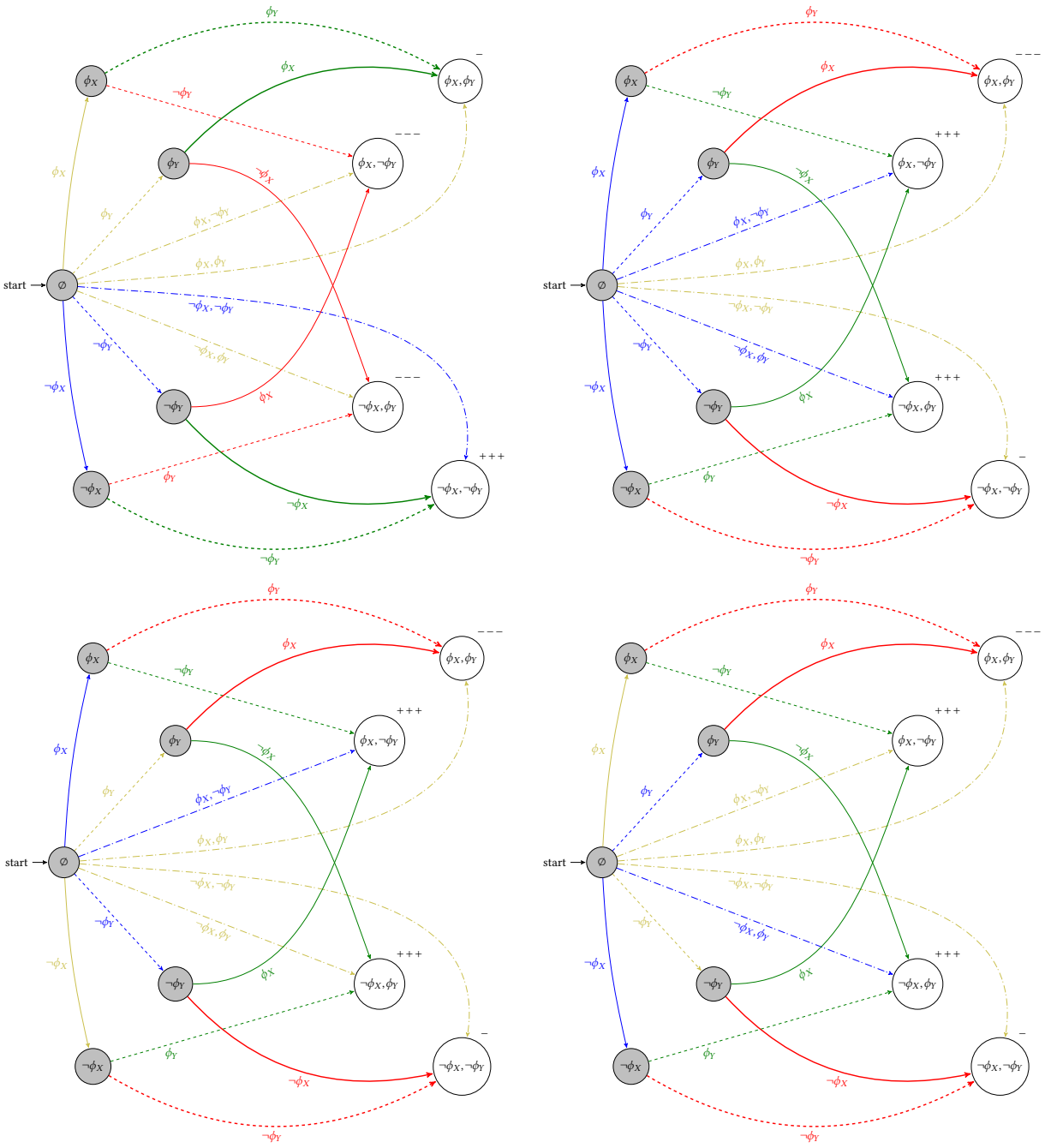


Figure 8.4: The GEFs of trouble (top left) and sailor (all other three GEFs). Green and red edges show assessments MOCOR already provides us with, even without any amendments (green edges correspond to right and red edges to wrong actions in the sense of MOCOR). Blue edges correspond to edges that should optimally be marked as right by MOCOR together with an amendment, yellow ones to edges such a version of MOCOR would *optimally* have assessed as wrong. In the last chapter, we have already found evidence that all still-considered amendments can deliver such optimal assessments for trouble-like cases. Even though it is not clear what should count as optimal assignments for sailor-like cases, all three plausible candidates in light of PMH are shown here. However, we cannot reasonably expect to find an optimal solution for these cases, anyway: while the top-right version contradicts our principle of how moral assessments of individual options translate to the moral status of combinations of simultaneous actions, the other two would violate Normative Supervenience (cf. Subsection 8.1.2). The question remains: what is the best performance camp MOAC can reasonably hope to achieve?

Requirements	Conceptual Deontic Consistency, Methodological Individualism, Being a MOAC Theory
Desiderata	Resolvability, No Moral Dilemmas, Weak Deontic Completeness, Deontic Completeness
Comparative Cachets	(Core) Extensional Adequacy, Parsimony, Simplicity

Table 8.2: A preliminary list of criteria as introduced in Chapter 4.

Fourth, further criteria discussed throughout the book are intended to serve as a central test criteria (cf. Table 8.2) for evaluating solutions, adhering to the principle of (Core) Extensional Adequacy (cf. Principle 4.3). This evaluation should not introduce unnecessary entities or complexity, respecting the principles of Parsimony and Simplicity.

Three candidate amendments remain to be considered: Summation, Maximization, and Mixed Strategies. First, we start by comparing Summation and Maximization. Let's call the policy that is induced by MOCOR together with Summation π^{Σ} and the policy that is induced by MOCOR together with Maximization $\pi^{\max}$.

First, we calculate the worth of π^{Σ}. For this amendment, I will perform the relevant deliberation in a detailed manner, applying the definitions and determinations from the preceding sections, starting with Definition 8.5applied to trouble. For other combinations of test cases and amendments, I will proceed less elaborately. Let us begin by recalling:

Definition 7.2 (Summation) *Let D be an individual decision situation with $\mathsf{D} = \langle A, \Phi, \mathsf{Out}_C : \Phi \to \mathcal{O} \rangle$ and actual context C. Further, let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be some valuation function. The* value of an individual decision situation *D is*

$$\mathsf{Val}^{\Sigma}(D) := \sum_{\phi \in \Phi} \mathsf{Val}_C(\phi).$$

Let D be a collective decision situation with actual context C and decomposition $\mathrm{I}_D = \{ \mathsf{D}_A \mid A \in \mathcal{A}_D \}$ and let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be some valuation function. The value of a collective decision situation *D is*

$$\mathsf{Val}^{\Sigma}(D) := \sum_{\mathsf{D}_A \in \mathrm{I}_D} \mathsf{Val}^{\Sigma}(\mathsf{D}_A).$$

To apply π^{Σ} to trouble, we first compute the following annotated normal form:

trouble		Y		$\mathsf{Val}^{\Sigma}_{\mathsf{trouble}}(\cdot)$
		ϕ_Y	$\neg\phi_Y$	
X	ϕ_X	$-\,(v_{11})$	$---\,(v_{12})$	$v_{11}+v_{12}$
	$\neg\phi_X$	$---\,(v_{21})$	$+++\,(v_{22})$	$v_{21}+v_{22}$
$\mathsf{Val}^{\Sigma}_{\mathsf{trouble}}(\cdot)$		$v_{11}+v_{21}$	$v_{12}+v_{22}$	

The Approach provides a decomposition of trouble, allowing us to identify the options for the agents' actions as selecting a row (or column, respectively). Accordingly, we have:

$$\begin{aligned}\mathsf{Val}^{\Sigma}(\phi_X) &= v_{11}+v_{12},\\ \mathsf{Val}^{\Sigma}(\neg\phi_X) &= v_{21}+v_{22},\\ \mathsf{Val}^{\Sigma}(\phi_Y) &= v_{11}+v_{21},\\ \mathsf{Val}^{\Sigma}(\neg\phi_Y) &= v_{12}+v_{22}.\end{aligned}$$

Since $v_{21}+v_{22} > v_{11}+v_{12}$ and $v_{12}+v_{22} > v_{11}+v_{21}$, we know that π^{Σ} will select $\neg\phi_X$ for X and $\neg\phi_Y$ for Y. From this, and using our tie-breaking assumption, i.e.,

$$\Pr^{T}_{D,C} : \Phi_D \to [0,1], \quad \Pr^{T}_{D}(\phi) = \begin{cases} \frac{1}{|T(D,C)|}, & \text{if } \phi \in T(D,C),\\ 0, & \text{otherwise,}\end{cases}$$

we directly infer the probabilities of the possible combinations of actions:

$$\pi^{\Sigma}(\langle\neg\phi_X,\neg\phi_Y\rangle) = \frac{1}{2} = \pi^{\Sigma}(\langle\neg\phi_Y,\neg\phi_X\rangle),$$

and $\pi^{\Sigma}(\Upsilon) = 0$ for all other combinations. The expected value of π^{Σ} for trouble is then:

$$\begin{aligned}\mathsf{Val}^{\Sigma}_{\mathsf{trouble}}(\pi^{\Sigma}) &= \frac{1}{2}\mathsf{Val}^{\Sigma}_{\mathsf{trouble}}(\langle\neg\phi_X,\neg\phi_Y\rangle) + \frac{1}{2}\mathsf{Val}^{\Sigma}_{\mathsf{trouble}}(\langle\neg\phi_Y,\neg\phi_X\rangle)\\ &= \frac{1}{2}v_{22} + \frac{1}{2}v_{22}\\ &= v_{22}.\end{aligned}$$

This is the (global) optimal outcome. Therefore, π^{Σ} not only closes the deontic gaps that gave rise to the Real Challenge but also successfully addresses the Challenge in classical Troublemakers, i.e., trouble.

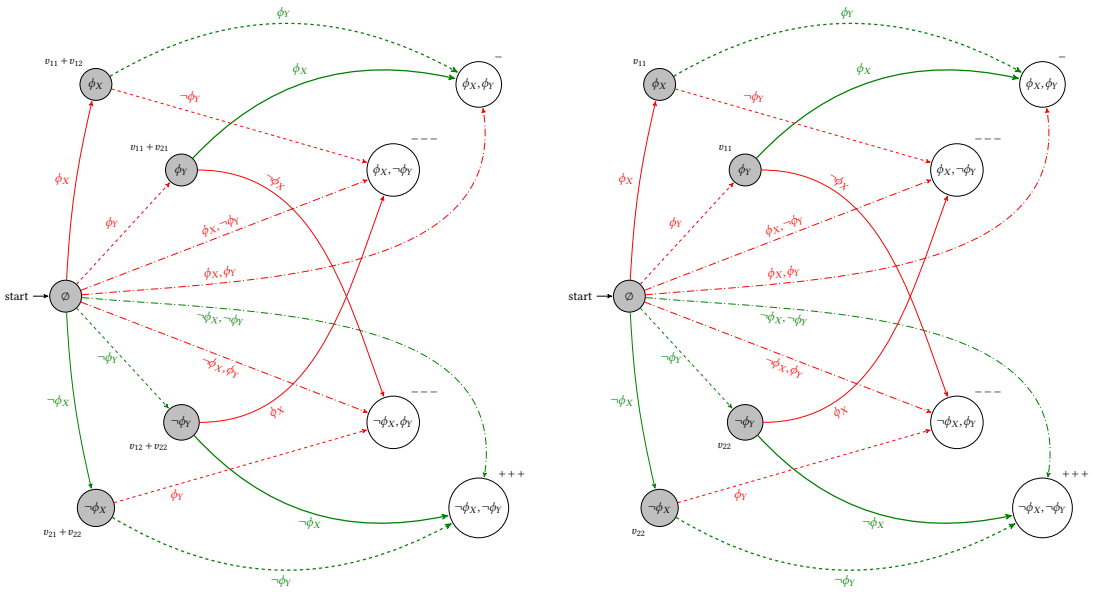


Figure 8.5: The annotated GEFs for trouble for MOCOR with Summation (left) and for MOCOR with Maximization. Both solutions are optimal (cf. Figure 8.4).

Next, we derive the (expected) value of $\pi^{\max}$ for trouble. As before, we start with an annotated normal form:

trouble		Y: ϕ_Y	Y: $\neg\phi_Y$	$\text{Val}^{\max}_{\text{trouble}}(\cdot)$
X	ϕ_X	$-$ (v_{11})	$---$ (v_{12})	v_{11}
	$\neg\phi_X$	$---$ (v_{21})	$+++$ (v_{22})	v_{22}
$\text{Val}^{\max}_{\text{trouble}}(\cdot)$		v_{11}	v_{22}	

Applying the same steps as in the case of π^{Σ}, we observe that $\pi^{\max}$ also achieves the optimal outcomes and, thus, has an expected value of v_{22}. Not only do both amendments yield the same results for traditional Troublemakers, but their policies also generate identical paths through the GEF of trouble (cf. Figure 8.5). These results are optimal: they successfully fill the normative gaps, address the REAL CHALLENGE, and enable MOAC to effectively master the CHALLENGE.

However, for sailor, the two amendments perform interestingly differently. Let us derive the expected values for the two policies relative to sailor, beginning with π^{Σ}. As before, we start by calculating the annotated normal form of sailor:

sailor		Y		$\mathsf{Val}^{\Sigma}_{\mathsf{sailor}}(\cdot)$
		ϕ_Y	$\neg\phi_Y$	
X	ϕ_X	$---\,(v_{11})$	$+++\,(v_{12})$	$v_{11}+v_{12}$
	$\neg\phi_X$	$+++\,(v_{21})$	$-\,(v_{22})$	$v_{21}+v_{22}$
$\mathsf{Val}^{\Sigma}_{\mathsf{sailor}}(\cdot)$		$v_{11}+v_{21}$	$v_{12}+v_{22}$	

As before, we compute:

$$\begin{aligned}\mathsf{Val}^{\Sigma}(\phi_X) &= v_{11}+v_{12},\\ \mathsf{Val}^{\Sigma}(\neg\phi_X) &= v_{21}+v_{22},\\ \mathsf{Val}^{\Sigma}(\phi_Y) &= v_{11}+v_{21},\\ \mathsf{Val}^{\Sigma}(\neg\phi_Y) &= v_{12}+v_{22}.\end{aligned}$$

Once again, we find $v_{21}+v_{22} > v_{11}+v_{12}$ and $v_{12}+v_{22} > v_{11}+v_{21}$, confirming that π^{Σ} selects $\neg\phi_X$ for X and $\neg\phi_Y$ for Y. Using the tie-breaking assumption, we derive the probabilities for the possible combinations of actions:

$$\pi^{\Sigma}(\langle\neg\phi_X,\neg\phi_Y\rangle) = \frac{1}{2} = \pi^{\Sigma}(\langle\neg\phi_Y,\neg\phi_X\rangle),$$

and $\pi^{\Sigma}(\Upsilon) = 0$ for all other combinations.
The expected value of π^{Σ} for sailor is thus:

$$\mathsf{Val}^{\Sigma}_{\mathsf{sailor}}(\pi^{\Sigma}) = \frac{1}{2}v_{22}+\frac{1}{2}v_{22} = v_{22}.$$

However, this time, π^{Σ} does not guide us to an optimal outcome. While it guarantees avoidance of the worst outcome (v_{11}), it leads us to a suboptimal result (v_{22}), falling short of the optimal outcomes v_{12} and v_{21}. Consequently, MOCOR amended by Summation ensures a degree of risk aversion but fails to achieve the best possible result for sailor.
Next, we compute the result for $\pi^{\max}$ for sailor, starting again with an annotated normal form:

sailor		Y		$\mathsf{Val}^{\max}_{\mathsf{sailor}}(\cdot)$
		ϕ_Y	$\neg\phi_Y$	
X	ϕ_X	$---\,(v_{11})$	$+++\,(v_{12})$	v_{12}
	$\neg\phi_X$	$+++\,(v_{21})$	$-\,(v_{22})$	v_{21}
$\mathsf{Val}^{\max}_{\mathsf{sailor}}(\cdot)$		v_{21}	v_{12}	

We can see immediately that the picture here is very different. The valuations of the options are:

$$\begin{aligned}\mathsf{Val}^{\max}(\phi_X) &= v_{12},\\ \mathsf{Val}^{\max}(\neg\phi_X) &= v_{21},\\ \mathsf{Val}^{\max}(\phi_Y) &= v_{21},\\ \mathsf{Val}^{\max}(\neg\phi_Y) &= v_{12}.\end{aligned}$$

Since $v_{11} = v_{21}$, and given our tie-breaker assumptions, we know that $\pi^{\max}$ will select all options with the same probability, namely with probability $1/2$. Accordingly, taking all orders of action into account, we see that $\pi^{\max}(\Upsilon) = 1/8$ for all combinations. The expected value of $\pi^{\max}$ for sailor is then:

$$\begin{aligned}\mathsf{Val}^{\max}_{\mathsf{sailor}}(\pi^{\max}) &= \frac{1}{8}\mathsf{Val}^{\max}_{\mathsf{sailor}}(\langle\phi_X,\phi_Y\rangle) + \frac{1}{8}\mathsf{Val}^{\max}_{\mathsf{sailor}}(\langle\phi_X,\neg\phi_Y\rangle)\\ &\quad+\frac{1}{8}\mathsf{Val}^{\max}_{\mathsf{sailor}}(\langle\neg\phi_X,\phi_Y\rangle) + \frac{1}{8}\mathsf{Val}^{\max}_{\mathsf{sailor}}(\langle\neg\phi_X,\neg\phi_Y\rangle)\\ &\quad+\frac{1}{8}\mathsf{Val}^{\max}_{\mathsf{sailor}}(\langle\phi_Y,\phi_X\rangle) + \frac{1}{8}\mathsf{Val}^{\max}_{\mathsf{sailor}}(\langle\phi_Y,\neg\phi_X\rangle)\\ &\quad+\frac{1}{8}\mathsf{Val}^{\max}_{\mathsf{sailor}}(\langle\neg\phi_Y,\phi_X\rangle) + \frac{1}{8}\mathsf{Val}^{\max}_{\mathsf{sailor}}(\langle\neg\phi_Y,\neg\phi_X\rangle)\\ &= \frac{1}{8}v_{11} + \frac{1}{8}v_{12} + \frac{1}{8}v_{21} + \frac{1}{8}v_{22}\\ &\quad+\frac{1}{8}v_{11} + \frac{1}{8}v_{12} + \frac{1}{8}v_{21} + \frac{1}{8}v_{22}\\ &= \frac{1}{4}v_{11} + \frac{1}{4}v_{12} + \frac{1}{4}v_{21} + \frac{1}{4}v_{22}\\ &= \frac{1}{4}v_{11} + \frac{1}{2}v_{12} + \frac{1}{4}v_{22}.\end{aligned}$$

Thus, the expected value of $\pi^{\max}$ for sailor is neither optimal, nor can it be ranked against the π^{Σ} independently from the concrete choice of values. For instance, for $v_{11} = -100$, $v_{22} = 0$, and $v_{12} = v_{21} = 100$, $\pi^{\max} > \pi^{\Sigma}$ (since $-25 + 50 = 25 > 0$), while if we had a case with $v_{12} = v_{21} = 20$, $\pi^{\max} < \pi^{\Sigma}$ (since $-25 + 10 = -15 < 0$).

Thus, we find that while MOCOR with Summation and MOCOR with Maximization perform significantly differently with respect to sailor, they both do not result in policies that achieve the optimal value as their expected value. While they cannot be ranked according to Definition 8.8, they exhibit different features of interest, though. For example, π^{Σ} *guarantees* that the worst outcome is avoided. In contrast, while $\pi^{\max}$ *cannot* guarantee this, it allows for the possibility of achieving one of the best outcomes, namely with a probability of $1/2$. However, as long as we have not examined the last candidate amendment, Mixed Strategies, we cannot yet determine whether these results ultimately speak against or in favor of Summation and Maximization.

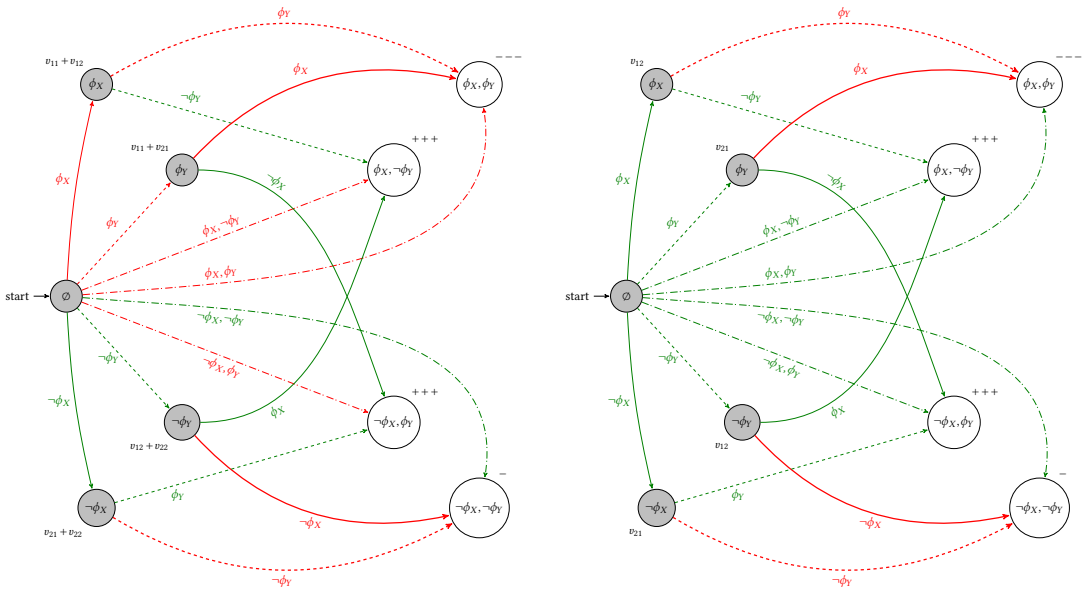


Figure 8.6: The annotated GEFs for sailor, one for MOCOR with Summation (left) and one for MOCOR with Maximization (right). Both solutions are *not* optimal (cf. Figure 8.4).

A closer examination reveals that, in addition, both policies exhibit notable issues with respect to the paths they permit for simultaneous actions. Interestingly, Summation proves to be too restrictive, while Maximization is overly permissive (cf. Figure 8.6, which illustrates these points). According to π^{Σ}, there are combinations of simultaneous actions containing wrong actions (i.e., red paths in the GEF) that still lead to optimal results. As a consequence, the resulting MOAC theory allows for scenarios where the best possible outcome is achieved, yet one agent acts wrongly. For instance, consider a case where Mikel gets the buoy at the very same moment Laika jumps, or vice versa; the optimal outcome is achieved, but one agent has acted wrongly. According to $\pi^{\max}$, all first actions are deemed right, which leads to the possibility that several combinations of simultaneous actions are right but produce suboptimal outcomes. For example, imagine both Mikel and Laika jump—or both fetch the buoy—at the very same moment. Such cases allow for suboptimal results despite all actions being considered right.[160]

While these objections are not conclusive—as they rely on the edge case of simultaneous action and are not surprising given that (a) the issue of simultaneous actions was intentionally left out when defining the expected value of policies, and (b) it has already been established that no amendment can be optimal (cf. Figure 8.4)—they nevertheless speak *pro tanto* against both amendments.

It is thus worth investigating the last remaining amendment, Mixed Strategies, to determine whether it can overcome these challenges while retaining

[160] A lexicographic version of $\pi^{\max}$ (cf. Footnote 144) would yield the same results as π^{Σ}.

the desirable features of Maximization and Summation. In the following, the program outlined by Smart in response to Brandt will be executed (cf. Subsubsection 7.3.2.3), in order to determine the expected value of π^{mix}.

We begin, as always, with an annotated normal form for trouble. While we have yet to determine the probability function according to which the two agents act, we know, in light of Normative Supervenience and the symmetry of our two test cases, that this distribution must be identical for both agents. Let us denote by p the shorthand for $\Pr(\phi_Z)$ (where $Z \in \{X, Y\}$). This gives us the following underspecified annotated normal form:

trouble		Y		$\mathsf{Val}^{\text{mix}}_{\text{trouble}}(\cdot)$
		ϕ_Y	$\neg\phi_Y$	
X	ϕ_X	$-\,(v_{11})$	$---\,(v_{12})$	$p \cdot v_{11} + (1-p) \cdot v_{12}$
	$\neg\phi_X$	$---\,(v_{21})$	$+++\,(v_{22})$	$p \cdot v_{21} + (1-p) \cdot v_{22}$
$\mathsf{Val}^{\text{mix}}_{\text{trouble}}(\cdot)$		$p \cdot v_{11} + (1-p) \cdot v_{21}$	$p \cdot v_{12} + (1-p) \cdot v_{22}$	

As mentioned earlier, we can determine p analytically using calculus. We start by explicating the function that describes the expected value as a function of the chosen probability distribution. Making use of the fact that $v_{12} = v_{21}$ in both types of test cases, we get:

$$\begin{aligned}
\text{EV}(p) &= p^2 v_{11} + p(1-p)v_{12} + p(1-p)v_{21} + (1-p)^2 v_{22} \\
&= p^2 v_{11} + 2p(1-p)v_{12} + (1-p)^2 v_{22} \\
&= p^2 v_{11} + 2pv_{12} - 2p^2 v_{12} + (1 - 2p + p^2)v_{22} \\
&= p^2 v_{11} + 2pv_{12} - 2p^2 v_{12} + v_{22} - 2pv_{22} + p^2 v_{22} \\
&= p^2(v_{11} - 2v_{12} + v_{22}) + 2p(v_{12} - v_{22}) + v_{22}.
\end{aligned}$$

We are interested in the maximum value for EV. Therefore, we compute the first derivative

$$\text{EV}'(p) = 2p(v_{11} - 2v_{12} + v_{22}) + 2(v_{12} - v_{22}).$$

Next, we set this first derivative equal to zero and solve for p:

$$\begin{aligned}
& & \text{EV}'(p) &= 0 \\
\Leftrightarrow & & 2p(v_{11} - 2v_{12} + v_{22}) + 2(v_{12} - v_{22}) &= 0 \\
\Leftrightarrow & & p(v_{11} - 2v_{12} + v_{22}) + v_{12} - v_{22} &= 0 \\
\Leftrightarrow & & p(v_{11} - 2v_{12} + v_{22}) &= v_{22} - v_{12},
\end{aligned}$$

thus:

$$p_{\text{extr}} = \frac{\overbrace{v_{22} - v_{12}}^{=:x}}{\underbrace{v_{11} - 2v_{12} + v_{22}}_{=:y}}.$$

We compute the boundary values and evaluate the extreme value:

$$\begin{aligned}
\mathrm{EV}(0) = v_{22}, \quad \text{and} \quad \mathrm{EV}(1) = v_{11},\\
\mathrm{EV}(p_{\mathrm{extr}}) &= \left(\frac{x}{y}\right)^2 \overbrace{(v_{11} - 2v_{12} + v_{22})}^{=y} + 2\frac{x}{y}\overbrace{(v_{12} - v_{22})}^{=-x} + v_{22}\\
&= \frac{x^2}{y} - 2\frac{x^2}{y} + v_{22}\\
&= v_{22} - \frac{x^2}{y}\\
&= v_{22} - \frac{(v_{22} - v_{12})^2}{v_{11} - 2v_{12} + v_{22}}.
\end{aligned}$$

Since $v_{11} < v_{22}$ for both test cases, and since v_{22} equals to the optimal solution of trouble, we can already conclude that $\mathsf{Val}^{\Sigma}_{\mathsf{trouble}}(\pi^{\mathrm{mix}}) = v_{22}$. Consequently, as expected at this point, MOCOR together with Mixed Strategies masters not only the REAL CHALLENGE but also Troublemakers, just as MOCOR together with Summation or Maximization did.

Next, we evaluate π^{mix} for sailor, comparing it to π^{Σ} and $\pi^{\max}$. Recall that sailor is characterized by:

$$v_{11} < v_{22} < v_{12} = v_{21}.$$

First we show that for sailor indeed $\mathrm{EV}(p_{\mathrm{extr}}) > v_{22}$. Rewriting sailor's characterizing order of outcome values above gives:

$$v_{11} - v_{12} < v_{22} - v_{12} < 0.$$

Checking for:

$$\begin{aligned}
& v_{22} - \frac{(v_{22} - v_{12})^2}{v_{11} - 2v_{12} + v_{22}} \overset{?}{>} v_{22}\\
\Leftrightarrow \quad & -\frac{(v_{22} - v_{12})^2}{v_{11} - 2v_{12} + v_{22}} > 0\\
\Leftrightarrow \quad & \frac{\overbrace{(v_{22} - v_{12})^2}^{>0}}{v_{11} - 2v_{12} + v_{22}} < 0\\
\Leftrightarrow \quad & v_{11} - 2v_{12} + v_{22} < 0\\
\Leftrightarrow \quad & v_{11} - v_{12} + v_{22} - v_{12} < 0.
\end{aligned}$$

Thanks to the above rewriting of the case-defining ordering, we know that

$$v_{11} - v_{12} < 0 \tag{8.1}$$

and that

$$v_{22} - v_{12} < 0. \tag{8.2}$$

but then obviously

$$v_{11} - v_{12} + v_{22} - v_{12} < 0.$$

Thus, we find that π^{mix} not only achieves an expected value greater than v_{22}—and therefore outperforms Summation on sailor—but also that p_{extr} is the relevant value to use in our valuation of π^{mix}. This leads to the following annotated normal form:

sailor		Y		$\text{Val}^{\text{mix}}_{\text{sailor}}(\cdot)$
		ϕ_Y	$\neg\phi_Y$	
X	ϕ_X	$---(v_{11})$	$+++(v_{12})$	$p_{\text{extr}} \cdot v_{11} + (1 - p_{\text{extr}}) \cdot v_{12}$
	$\neg\phi_X$	$+++(v_{21})$	$-(v_{22})$	$p_{\text{extr}} \cdot v_{21} + (1 - p_{\text{extr}}) \cdot v_{22}$
$\text{Val}^{\text{mix}}_{\text{sailor}}(\cdot)$		$p_{\text{extr}} \cdot v_{11} + (1 - p_{\text{extr}}) \cdot v_{21}$	$p_{\text{extr}} \cdot v_{12} + (1 - p_{\text{extr}}) \cdot v_{22}$	

Now, we already know that the expected value of π^{mix} for sailor is (just as for all minimal Coordination Cases described with these four variables for the values of its final outcomes):

$$\text{Val}^{\text{mix}}_{\text{sailor}}(\pi^{\text{mix}}) = v_{22} - \frac{(v_{22} - v_{12})^2}{v_{11} - 2v_{12} + v_{22}}.$$

The final question now is whether π^{mix} can outperform π^{max} on sailor, i.e., whether:

$$\text{Val}^{\text{mix}}_{\text{sailor}}(\pi^{\text{mix}}) > \text{Val}^{\text{max}}_{\text{sailor}}(\pi^{\text{max}}),$$

and thus:

$$v_{22} - \frac{(v_{22} - v_{12})^2}{v_{11} - 2v_{12} + v_{22}} > \frac{1}{4}v_{11} + \frac{1}{2}v_{12} + \frac{1}{4}v_{22}.$$

The answer is clear: yes! This is evident because, even if choosing with $p = \frac{1}{2}$—and thus randomizing equally between actions, as π^{max} yields for sailor—were optimal, then p_{extr} would, by construction, precisely equal $\frac{1}{2}$. Therefore, π^{mix} inherently captures the optimal probability distribution, ensuring that it outperforms or at least matches any alternative probability assignments, including the uniform distribution.

In conclusion, π^{mix}, while mastering the Real Challenge and simultaneously solving the Challenge for trouble, also surpasses π^{Σ} and, in nearly all plausible value configurations, π^{max} on sailor. Things look promising for this amendment.

The results of this section, summarized in Table 8.3, pave the way for a final judgment on the amendments considered in this work.

Amendment	Best *expected value* for all cases	*Guarantees* best outcome for classical Troublemakers	*Guarantees* best outcome for Seaman Clumsy–like cases	*Could* lead to best outcome in Seaman Clumsy–like cases	*Guaranteed* not worst outcome in Seaman Clumsy–like cases	Best EV for Seaman Clumsy–like cases	further pros	further cons
Summation	×	✓	×	×	✓	×	simple	assesses some optimal paths as wrong, thus assesses *combinations* rather too restrictively
Maximization	×	✓	×	✓	×	×	connects well with the "M" in "MOAC", all optimal paths are assessed as right	assesses *some* combinations too permissible
Maximization (lexicographic)	×	✓	×	×	✓	×	connects well with the "M" in "MOAC"	assesses some optimal paths as wrong, thus assesses *combinations* rather too restrictively
Mixed Strategies	✓	✓	×	✓	×	✓	based on a well-established concept from instrumental rationality	requires adding a new formal methodology to the consequentialist toolbox

Table 8.3: All three candidate amendments cope perfectly with Troublemakers. Mixed Strategies copes best with Seaman Clumsy–like cases, but also comes with the price that the catastrophic result may occur. Summation and lexicographic Maximization lead to qualitatively identical outcomes.

8.3.2 And the Winner Is ...

The results are *not* conclusive regarding the overall best amendment. Classical Troublemakers (i.e., test case trouble) can be handled optimally by all three amendments, while Seaman Clumsy–like cases (i.e., test case sailor) cannot be solved optimally by any of the three. This limitation, however, aligns with the 'reasonable impossibility results' discussed in Section 8.1, and thus cannot be held against the amendments. Still, the amendments perform significantly differently on sailor, even though Maximization is arguably subsumed by Mixed Strategies.

Ultimately, then, the final choice hinges on one's tolerance for risk. Where Mixed Strategies allows for the possibility of achieving the optimal outcome in Seaman Clumsy, it also introduces the risk of the worst-case scenario. On the other hand, Summation guarantees avoiding the worst outcome but does so at the cost of excluding the possibility of achieving the best in sailor—for example, successfully rescuing Clumsy.

An example with concrete values is helpful for illustration. Let us assume that the death of all three sailors results in a disvalue of −9000. Conversely, rescuing Clumsy would have a value of 3000. A drowning of Clumsy would have a disvalue of −3000. This setup implies the following instance of Seaman Clumsy:

		Laika	
		jump	get buoy
Mikel	jump	−9000	+3000
	get buoy	+3000	−3000

We have seen that naïve Maximization allows for all possible combinations of actions. So, given that the agents choose by fair randomization between the right actions, this would mean a 25% chance of every outcome, resulting in an expected value of

$$\mathrm{EV}^{\max}_{\mathsf{sailor}} = \frac{1}{4}(-9000) + \frac{1}{4}(3000) + \frac{1}{4}(3000) + \frac{1}{4}(-3000) = -1500.$$

Lexicographic Maximization and Summation both result in the rightness of getting the buoy for both agents and, thus, in a certain disvalue of −3000.[161]

[161] Note that lexicographic Maximization and Summation can also outperform naïve Maximization for certain values. For instance, consider a variant where the drowning of Clumsy would have a disvalue of rather −1000.

For Mixed Strategies, the probability for jumping is given by the general formula derived earlier:

$$p = \frac{\overbrace{-3000 - 3000}^{v_{22} - v_{12}}}{\underbrace{-9000 - 2 \cdot 3000 - 3000}_{v_{11} - 2v_{12} + v_{22}}} = \frac{-6000}{-18000} = \frac{1}{3}.$$

For the expected value, we calculate:

$$\begin{aligned}
\mathrm{EV}^{\mathrm{mix}}_{\mathrm{sailor}} &= \overbrace{-3000}^{=v_{22}} - \frac{\overbrace{(-3000 - 3000)^2}^{=(v_{22} - v_{12})^2}}{\underbrace{-9000 - 2 \cdot 3000 - 3000}_{=v_{11} - 2v_{12} + v_{22}}} \\
\mathrm{EV}^{\mathrm{mix}}_{\mathrm{sailor}} &= -3000 - \frac{(-6000)^2}{-18000} \\
&= -3000 - \frac{36000000}{-18000} \\
&= -3000 + 2000 \\
&= -1000.
\end{aligned}$$

As expected, Mixed Strategies achieves a better expected value than both Maximization and Summation, while also allowing for the possibility of achieving the best outcome, i.e., rescuing Clumsy, with a $\frac{4}{9}$ probability. However, there is also a $\frac{1}{9}$ chance of ending up with the worst outcome, where all jump and die, as well as a $\frac{4}{9}$ chance of arriving at the second-worst outcome, where Mikel and Laika both get the buoy and, in the meantime, Clumsy drowns.

So, what is the final verdict? Ultimately, the usual arguments for Mixed Strategies align with those supporting expected utility calculations in the context of rational decision-making under risk. In particular, when applied to the totality of all (or even 'just' to the totality of the actually occurring) situations, Mixed Strategies will pay off morally in aggregate.[162] And the aggregate of the moral good is what matters most for camp MOAC.

I therefore advocate for Mixed Strategies as the collective amendment of choice for MOAC. In my view, thus, it is justified to modify MOCOR to incorporate the principle of maximizing expected value (cf. Subsubsection 7.3.2.3). Accordingly, I propose replacing MOCOR with Mixed Strategies, now reintroduced under the name *Multi-Agent Maximizing Objective Criterion of Rightness* (MA-MOCOR):

[162] This is essentially a classical long-run argument for expected utility as the foundation for rational choice, cf. Briggs 2023, sec. 2.1.

Principle 8.2 (MA-MOCOR) *Let D be a collective decision situation with $\mathsf{D} = \langle \mathcal{A}, \Gamma, \mathsf{Out}_C : \Psi_\Gamma \to \mathcal{O} \rangle$, where $A \in \mathcal{A}$ is an agent, $\phi \in \Phi_A \in \Gamma$ is an option of that agent A, and C is the actual context. Additionally, let $\mathsf{Val} : \mathcal{W} \to \mathbb{R}$ be a valuation function. To perform an action $\phi \in \Phi_A$ in D with probability $\Pr_{\Phi_A}(\phi)$ is* right for *A in C if and only if A's choice follows a probability distribution $\Pr_{\Phi_A}$ that belongs to a set of acceptable probability distributions for all agents, such that $\Pr_\Psi$, given by*

$$\Pr_\Psi(\Upsilon) := \prod_{A \in \mathcal{A}} \Pr_{\Phi_A}(\phi_i),$$

maximizes the expected value of the outcomes relative to C, expressed as:

$$\mathrm{EV}(\Pr_\Psi) := \sum_{\Upsilon \in \Psi_\Gamma} \Pr_\Psi(\Upsilon) \cdot \mathsf{Val}_C(\Upsilon),.$$

I shall call the family of theories subscribing to this criterion of rightness Multi-Agent Consequentialism (MAC) theories. MAC theories are fit for multi-agent scenarios, and it must be noted that for 'normal' decision situations (i.e., individual decision situations where actions immediately result in final outcomes), MOCOR and MOCOR (EU) are extensionally equivalent. Therefore, first, for 'traditional' individual decision situations, this theoretical adjustment cannot lead to new challenges; and second, MAC should be understood as a *generalization* of MOAC.

Of course, Mixed Strategies comes with some conceptual challenges. On the one hand, one must accept that not only the explicitly available actions exist but that it is at least in principle possible for agents to act according to arbitrary probability distributions. On the other hand, one must accept that acting according to such probability distributions now becomes the default for doing the right thing. However, I do not think this should be taken as a specific challenge to my approach. In fact, Mixed Strategies is in the best of company. For instance, in instrumental rationality, it is commonplace to accept mixed strategies, and dismissing probabilistic approaches to the selection of actions would result in massive losses in expected value and, thus, in the performance of our best theories of instrumental rationality. Therefore, I see no proper reason to reject Mixed Strategies as an amendment of choice for camp MOAC. Consequently, I argue for the acceptance of Mixed Strategies as an amendment in the sense of reasonable MH to solve the REAL CHALLENGE without introducing a new variant of CHALLENGE.

Before I retrace the success of my approach in the context of an overall conclusion, allow me to make a comment concerning the non-exhaustiveness of my investigation: Of course, it cannot be excluded that the consideration of further amendments against the background of other, possibly more extensive and complex kinds of collective decision situations could turn out to be

advantageous. Further, it cannot be excluded that different amendments for different (classes of) collective decision situations could prove beneficial. In such cases, camp MOAC should consider an *ensemble solution*, i.e., incorporate several amendments that are used singly or in combination depending on the structure of the specific collective decision situation to be assessed. Ensemble solutions have already shown promise in formal decision theory, especially in the area of machine learning (cf. Dong et al. 2020). A family of theories that has been unafraid of genuine conditional assessment for so long (recall the CSM) is unlikely to shy away from the conditional use of amendments.

8.4 On Overall Success

In light of Chapter 6, the REAL CHALLENGE had to be addressed first before we could even return to the Pyramid and, by extension, to the CHALLENGE in all its complexity (cf. Figure 8.7). We recall the related argument for the REAL CHALLENGE:

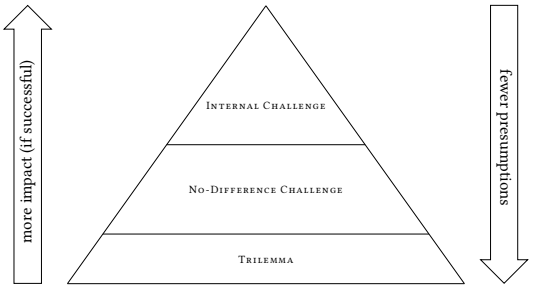


Figure 8.7: The three variants of the CHALLENGE, shown as the Pyramid, organized according to their potential impact and their presumptions.

The Real Argument

$P_{\exists \text{NON-DEC}}$: There are non-decomposable collective decision situations, i.e., collective decision situations in which, relative to the actual and normatively sufficiently complete context, it is undefined for at least one agent, with respect to at least two of their options, how these options are to be ranked based on the moral quality of their actual consequences.

$P_{\exists \text{GAPS}}$: If there are such collective decision situations, then there is a broad class of decision situations for which MOAC offers no meaningful guidance; in other words, it results in (a myriad of) systematic deontic gaps.

$P_{\text{NO GAPS!}}$: If a moral theory is adequate, then there is no broad class of decision situations for which MOAC offers no meaningful guidance; in other words, it must not result in (a myriad of) systematic deontic gaps.

$C_{\neg \text{ADEQ}}$: MOAC is not adequate.

Thanks to the APPROACH, $P_{\exists \text{NON-DEC}}$ has long been resolved, and Compositionalism has proven to be defensible. Each action of each agent has a well-defined consequence: the decision situation that remains for all other agents, contingent upon that agent's action. However, while the discovery of these 'new' consequences offers significant progress, it is not sufficient by itself. As long as we cannot rank these newly discovered consequences, a multitude of actions remain unrankable with respect to the moral qualities of their outcomes, leaving normative gaps unresolved. Only by adding a collective amendment, these gaps could be bridged, enabling camp MOAC to effectively master the REAL CHALLENGE.

This naturally raised the question of how these gaps *should* be closed—specifically, which collective amendment should be chosen. Given what we had already learned about the CHALLENGE in the first part of this project, it was evident that any amendment selected must not simply reintroduce the CHALLENGE. With the Pyramid in mind, our focus naturally centered on addressing the CHALLENGE as INTERNAL CHALLENGE and, consequently, on PMH. We recall:

The ARGUMENT – tentative

$P_{\exists \text{TROUBLE}}$: There are Troublemakers: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

P_{MOCOR}: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

P_{MH}: If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (according to this theory).

$C_{\neg \text{ADEQ}}$: MOAC is not an adequate moral theory.

Given a collective amendment and the consequences revealed by the APPROACH, there is a plausible interpretation of P_{MOCOR} that avoids the peculiar ex post reasoning that originally undermined the validity of the ARGUMENT. The task, then, was to identify an amendment that would allow camp MOAC to condemn at least one action in every combination leading to suboptimal

results—thereby ensuring that the amendment does not violate the condition specified on the rightmost side of P_{MH}.

Before fully endorsing MH (and thereby accepting P_{MH}) and directing all efforts towards it, it was crucial to revisit PMH for further scrutiny. In fact, our examination of Seaman Clumsy (in Section 8.1) uncovered a non-trivial counterexample to MH. Somewhat surprisingly, we found that a theory satisfying MH *cannot* also be an objectively consequentialist theory. In Seaman Clumsy–like cases, such a theory would necessarily violate the principle of Normative Supervenience, a principle inherent in the very definition of objective consequentialism.

Ultimately, this led us to refine MH by introducing a restriction to the *best reasonably expectable results*. This refinement was successfully articulated through the concepts of *theory-induced policies* and their *expected value relative to* (sets of) *decision situations*. Through this framework, a clear winner emerged: Mixed Strategies—the collective amendment of choice for MOAC.

In this respect, the CHALLENGE can also be regarded as mastered in the form of the INTERNAL CHALLENGE. We recall the corresponding solution space (see Figure 8.8 on the next page). The amendment of choice (cf. Subsection 8.2.2), Mixed Strategies, defines a modified version of MOCOR, referred to as MA-MOCOR, ultimately giving rise to Multi-Agent Consequentialism (MAC). In traditional 'Hi-Lo' Troublemakers such as Two Factories or Whiff and Poof, MAC is guaranteed identifies wrong actions in combinations that result in suboptimal outcomes and recommends only actions that, if followed, inevitably lead to one of the best possible outcomes.

For more demanding cases like Seaman Clumsy, which feature multiple optima not located on the main diagonal, MAC may, in rare instances, lead to the worst possible result. Such an outcome, however, would be attributable to bad moral luck—a trade-off worth accepting, as this amendment arguably *maximizes the total expected value*, at least in the long run. Consistently applied across all (possible or actual) decision situations of the investigated types, MAC thus ensures the greatest aggregate moral good.

Note that in terms of the solution space for the INTERNAL CHALLENGE, MAC, strictly speaking, takes two exits simultaneously: both P_{MOCOR} and P_{MH} are falsified, each due to the unwarranted assumption of "necessity". However, when restricted to Troublemakers, MAC only 'falsifies' P_{MOCOR}.

This resolution significantly aids progress when addressing the second level of the Pyramid: the CHALLENGE as NO-DIFFERENCE CHALLENGE. We recall the associated solution space (see Figure 8.9, also on the next page) and the corresponding argument:

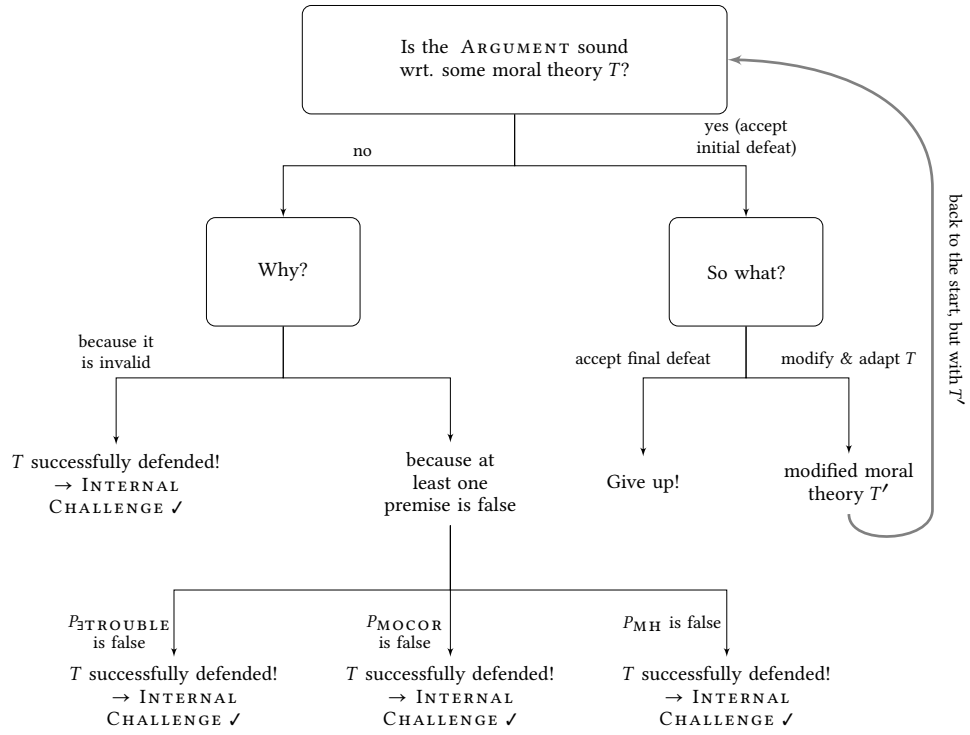


Figure 8.8: The solution space of the Challenge as Internal Challenge.

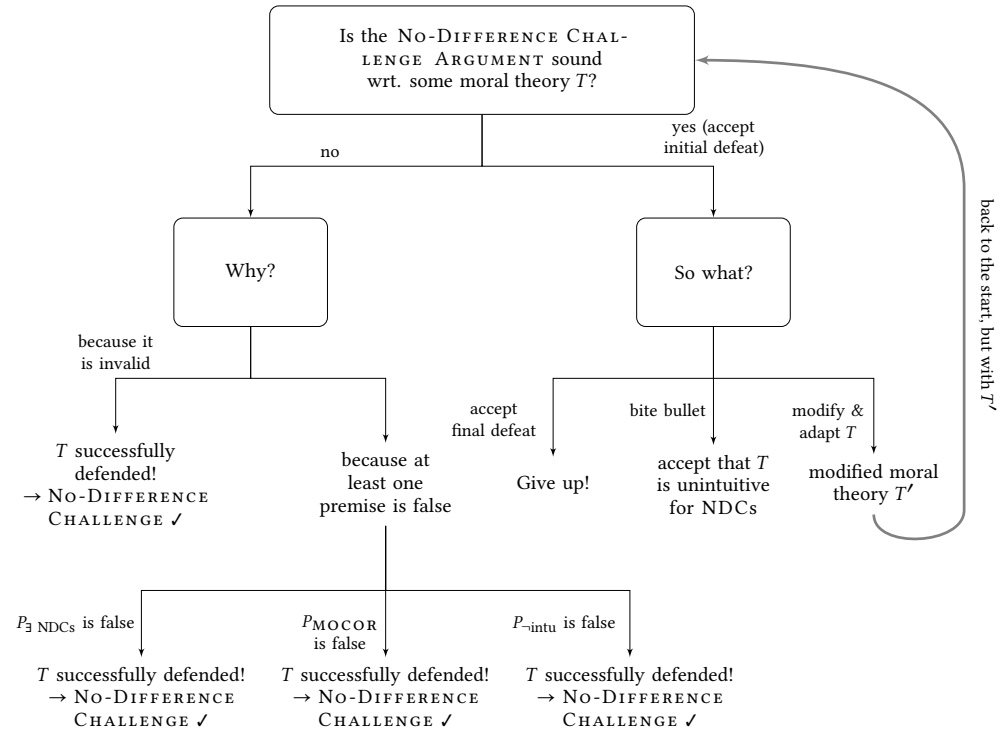


Figure 8.9: The solution space of the Challenge as No-Difference Challenge.

The No-Difference Challenge Argument – tentative

$P_{\exists \text{NDCs}}$: There are No-Difference Cases: collective decision situations in which there is at least one agent who can act in a way such that it seems intuitively morally wrong, but the agent could not make a difference for the morally better by unilaterally acting differently.

P_{MOCOR}: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

$P_{\neg\text{intu}}$: If a moral theory assesses intuitively morally wrong actions as morally right for a significant class of situations, then T is counterintuitive.

$C_{\neg\text{intu}}$: MOAC is counterintuitive.

Again, P_{MOCOR} is false according to MAC, and so the No-Difference Challenge is solved automatically.

However, MAC may conflict with certain intuitions, including those inspired by MOAC itself. For instance, its recommendation to act with specific probabilities in cases like Seaman Clumsy might strike some as counterintuitive. Indeed, when presenting my approach in recent years, I have occasionally encountered incredulous reactions for this suggestion. Yet, I believe that 'informed and well-considered' intuitions—particularly those aligned with consequentialist thinking—should find this approach reasonable. For camp MOAC, the solution lies in following well-established principles of formal decision and game theory and embracing the possibilities, as well as the power, offered by probabilistic strategies.

After all, consequentialists must *somehow* choose between multiple right options, even if cases involving two exactly morally equivalent consequences feel unfamiliar or exotic. If one finds the probabilistic approach unpalatable and insists on not biting the probabilistic bullet, alternatives like Summation or the naïve or lexicographic versions of Maximization remain available. While these alternatives generally forgo some expected value in Seaman Clumsy–like cases, they avoid the conceptual challenge of endorsing probabilistic actions. Although I see no compelling reason to reject the probabilistic approach, adopting one of these alternatives might make the choice more acceptable to certain proponents of camp MOAC in No-Difference Cases.

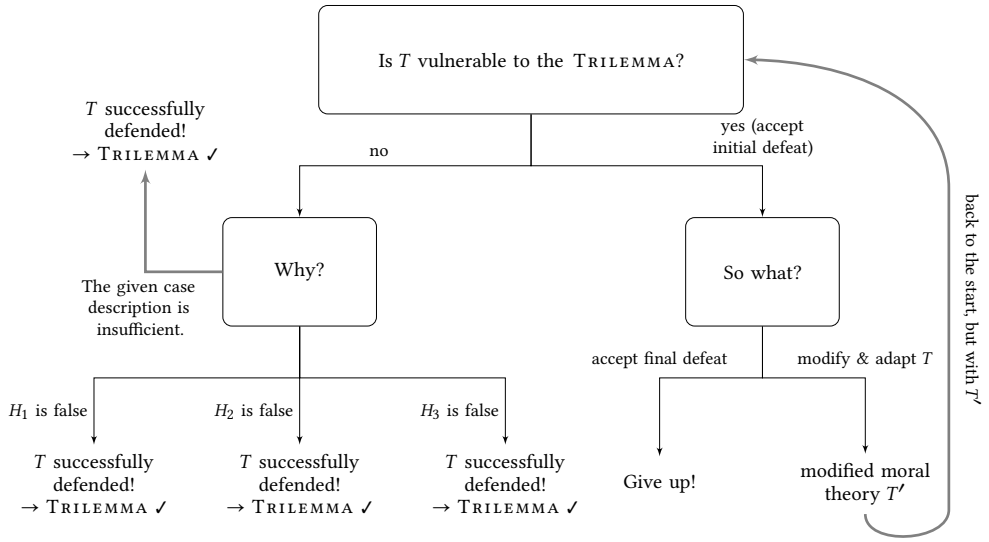


Figure 8.10: The solution space of the Trilemma.

In the end, this only leaves us with the Trilemma. We recall its solution space (see Figure 8.10). Solving the Trilemma required rejecting one of the following three propositions :

(H_1) What Ann and Ben have done is morally wrong.

(H_2) If something morally wrong has been done, then someone did wrong.

(H_3) Neither Ann nor Ben did wrong.

MAC has no problem with rejecting H_3. The first acting agent did wrong—and if both acted simultaneously, both acted wrongly.

And with that, we have come to the end of a long journey, dug through the Pyramid, and can now say: MOAC in the form of MAC has mastered the Challenge, in its strongest form (cf. Figure 8.11).

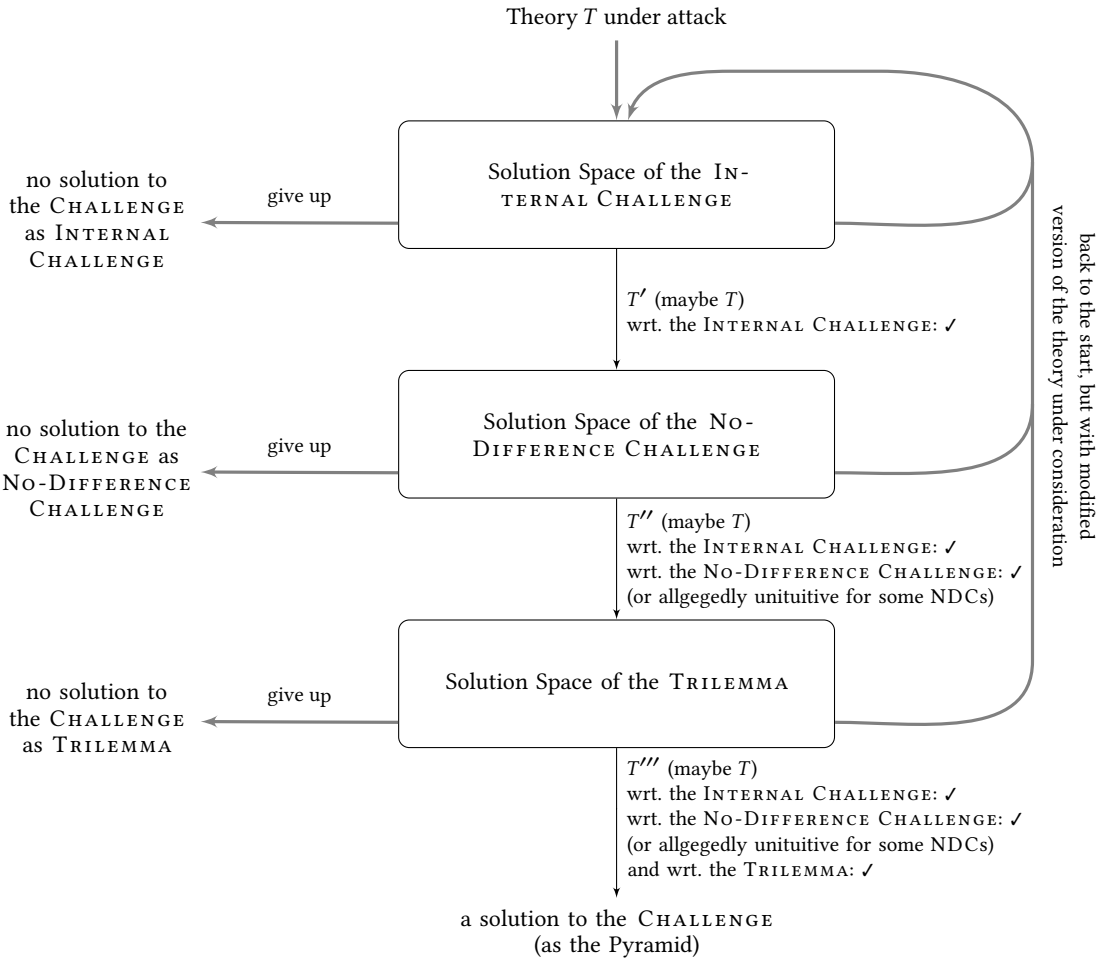


Figure 8.11: The solution space of the Challenge as the Pyramid. We have reached the bottom, which is actually a climax.

Conclusion

Chapter 9

Summary and Future Work

Central to this undertaking was an exploration and resolution of the Challenge of Collective Action (in short, the Challenge) through the framework of Maximizing Objective Act-Consequentialism (or simply moac). I claim to have succeeded—against all odds. As the project unfolded, its scope expanded beyond my initial expectations, veering into more technical and formal territory than I had originally anticipated. Reflecting on the overall effort, I believe at least four philosophical insights deserve mention.

Firstly, the Challenge is best understood not as a single challenge but as a mosaic of interconnected intricacies. Its various manifestations arise from the apparent disparity between expected and actual moral assessments in specific collective decision situations, particularly Troublemakers. Each variant of the Challenge introduces distinct implications regarding the potential consequences of success, while also differing in their preconditions. Thus, the Challenge should be seen as a collection of mutually supporting challenges, where some aspects reinforce others. My reconstruction of this multi-leveled version, represented as the Pyramid, comprises three levels: the Challenge as Internal Challenge, the Challenge as No-Difference Challenge, and the Challenge as Trilemma (cf. Figure 9.1).

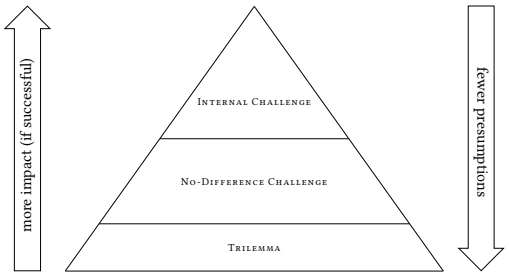


Figure 9.1: The three variants of the Challenge, shown as the Pyramid, organized according to their potential impact and their presumptions. The higher a variant in this hierarchy, the greater the potential impact on moac. Conversely, a variant further down in the hierarchy requires fewer presumptions to be successful. All three variants were successfully addressed in this project.

Secondly, a deeper examination reveals that Internal Challenge—the most challenging yet also the most assumption-laden variant of the Challenge—does not hold up under scrutiny and must ultimately be dismissed as invalid. This variant has drawn attention primarily because it underscores the implications of a reasoning strategy that camp moac adopted to avoid

outright failure in collective decision situations. However, a closer investigation of MOAC reveals a pervasive moral silence in numerous collective contexts, where MOCOR at best enables conditional assessments. This is the REAL CHALLENGE, the root cause of the collective struggle faced by MOAC theories. In an effort to bridge these moral voids and address the resulting deontic gaps at any cost, consequentialists have 'enriched' the contexts in collective decision-making to include actions across all agents. Refuting the strongest form of the CHALLENGE, therefore, is not—as such—a victory for consequentialism. Rather, it highlights that camp MOAC has sidestepped a core issue and an even more fundamental challenge: the REAL CHALLENGE.

Fortunately, the third essential finding of my thesis offers a pathway for meaningful progress. It is indeed possible to precisely define appropriate consequences for each action taken by individual agents within any collective decision situation. This only requires a willingness to recognize decision situations as potential (intermediate) outcomes, in line with established conventions in other decision-theoretic paradigms. This Intermediate Outcomes Approach (or simply, the APPROACH) enables the decomposition of arbitrary collective decision situations into individual decision situations, which can then be systematically evaluated by MOCOR. By assessing these newly identified consequences with the help of so-called collective amendments, such as Mixed Strategies, camp MOAC is equipped to bridge deontic gaps in a principled and less *ad hoc* manner, thereby addressing the REAL CHALLENGE without reigniting the CHALLENGE.

Fourthly, and finally, we revisited the Principle of Moral Harmony (or PMH), which asserts that when multiple agents act rightly in unison, they are guaranteed to bring about the morally best outcome they can collectively achieve. However, critical reflection revealed that existing formulations of PMH were overly rigid. These definitions, in their traditionally broad scope, conflict both logically and conceptually with the foundational principles of objective consequentialism. To resolve this tension, I introduced a tempered alternative: reasonable Moral Harmony (or simply, reasonable MH). Guided by this principle, I explored various collective amendments to leverage the newly identified consequences. My investigation culminated in the adoption of Mixed Strategies as a theoretical amendment to the MOAC framework, leading to the evolution of MOCOR into MA-MOCOR, and ultimately giving rise to Multi-Agent Consequentialism (MAC).

Throughout this research, I have introduced new tools and methodologies to advance the consequentialist framework. Most notably, the generalized extensive form (GEF) offers a structured representation of collective decision situations, while the concept of theory-induced policies, provides a principled method for evaluating moral theories based on their performance within such GEFs. These theoretical innovations not only strengthen the analytical

foundation of camp MOAC but also enhance its capacity to address collective challenges—and potentially extend to other areas of inquiry.

This progress not only resolves key challenges within the current framework but also lays a solid foundation for future exploration and refinement, paving the way for further advancements.

9.1 Future Work?

While this thesis provides an in-depth exploration, it doesn't claim to be exhaustive. Various ideas, examples, and arguments were shelved, either due to deliberate limitations regarding the scope of this endeavor or due to their emergence late in the research process. What follows is an overview of potential avenues for future investigation, building on the foundation established in this work.

9.1.1 Implications for Subjective Consequentialism

There are, of course, varying perspectives on the relationship between objective and subjective variants of consequentialism. As discussed in Subsection 2.3.1, particularly in the context of Principle 2.4, these variants can be viewed either as opposites or as complementary approaches. I lean toward the latter view, holding that subjective consequentialist theories should aim to develop decision procedures and, potentially, criteria of rightness 'for real agents' that *approximate* the 'true' objective criterion of rightness. Whether or not one shares this stance, it is an established perspective with significant implications derived from this work.

In this thesis, I have articulated what is right according to MOAC (modulo axiological background theories) and outlined the methods for determining this. Consequently, subjective theories—understood as relaxations or approximations of the objective rightness predicate—should be expected to adapt to these foundations. Exploring these implications further is a natural next step, and I have identified this endeavor as my next normative ethical research project. This investigation will involve parameterizing subjective theories to accommodate different types of agents, including artificial ones. I believe this direction will lay the groundwork for a research program in machine ethics that is deeply rooted in robust philosophical theory.

9.1.2 Implications for the Actualism–Possibilism Debate

Recall the following part of A. N. Prior's quote from Section 6.5 (Prior and Raphael 1956, pp. 91–92):

> Suppose that determinism is *not* true. Then there may indeed be a number of alternative actions which we could perform on a given occasion, but none of these actions can be said to have any "total consequences", or to bring about a definite state of the world which is better than any other that might be brought about by other choices. For we may presume that other agents are free beside the one who is on the given occasion deciding what he ought to do, and the total future state of the world depends on how these others choose as well as on how the given person chooses; and even if there were not other people to spoil one's calculations there would still be oneself, with one's own future choices, or some of them, undetermined like this present one (unless a man decides that it is too risky for him to have any further freewill, and on this very ground finds it to be his duty to do away with himself).

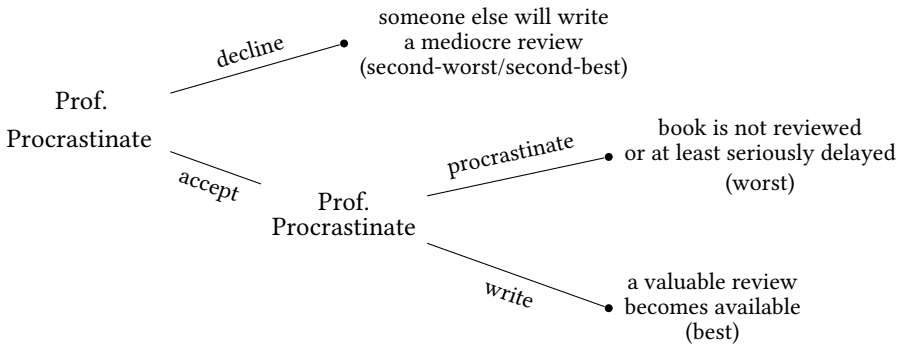
Prior undoubtedly raises an important point by emphasizing that other agents are not strictly necessary to raise many of the challenges discussed in this work. Indeed, an analogous issue, where the occurrence of certain consequences depends on what the same agent does at a later point in time, has long been a central topic in the actualism–possibilism debate. This debate, which has been mentioned several times in this project, dates back at least to the late 1960s, when Hector-Neri Casta neda (1968) observed that MOAC theories often struggle to adequately assess the rightness of individual actions, even if they can evaluate entire sequences of actions according to their own criteria of rightness.[163] Much like in the case of Troublemakers, it is trivial for camp MOAC to identify the best possible combinations of actions, yet mapping this assessment appropriately to individual actions proves difficult, if not impossible. Structurally, this is the same issue, even though the agents (or, maybe more precisely, the temporal parts) involved are more intimately related. The *locus classicus* of modern discussions on this topic remains the article by Frank Jackson and Robert Pargetter (Jackson and Pargetter 1986), which includes the following now-famous example (ibid., p. 235):

Case 9.1 (Professor Procrastinate) *Professor Procrastinate receives an invitation to review a book. He is the best person to do the review, has the time, and so on. The best thing that can happen is that he says yes, and then writes the review when the book arrives. However, suppose it is further the case that were Procrastinate to say yes, he would not in fact get around to writing the review. Not because of incapacity or outside interference or anything like that, but because he would keep on putting the task off. (This has been known to happen.) Thus, although the best that can happen is for Procrastinate to say yes and then write, and he can do exactly this, what would in fact happen were he*

[163] Chrisoula Andreou, in her exploration of the CHALLENGE (Andreou 2014), which she has noted in private correspondence was never intended for consequentialists, seems to address a similar observation.

to say yes is that he would not write the review. Moreover, we may suppose, this latter is the worst that can happen. It would lead to the book not being reviewed at all, or at least to a review being seriously delayed.

Of course, this case can be represented in an extensive form:



It is clear that Professor Procrastinate warrants conditional assessments analogous to those encountered in the context of the CHALLENGE. For instance:

(50) If Prof. Procrastinate accepts the invitation to review the book, it is right for him to write it (and wrong for him to procrastinate).

(51) If Prof. Procrastinate will procrastinate, it is right for him to decline the invitation to review the book (and wrong for him to accept it).

While Professor Procrastinate shares a structural similarity with Job Market, thereby resembling a Threshold Case, there are also scenarios that appear to lack a temporal component, making them more akin to Coordination Cases. One such example is presented by Holly Goldman (1978, p. 186), which is also referenced by Jackson and Pargetter:

Case 9.2 (JONES) *Jones is driving through a tunnel behind a slow-moving truck. It is illegal to change lanes in the tunnel, and Jones's doing so would disrupt the traffic. Nevertheless, she is going to change lanes—perhaps she doesn't realize it is illegal, or perhaps she is simply in a hurry. If she changes lanes without accelerating, traffic will be disrupted more severely than if she accelerates. If she accelerates without changing lanes, her car will collide with the back of a truck.*

This decision situation can arguably be represented quite adequately in the following (in this case likely asymmetric) normal form:

JONES	accelerate	maintain speed
change lane	−	−−
stay on lane	−−−	+

As before with Professor Procrastinate, JONES implies conditional assessments. For example:

(52) If Jones changes lanes, it is right for him to accelerate (and wrong for him to maintain speed).

(53) If Jones maintains speed, it is right for him to stay on the lane (and wrong for him to change lanes).

Unlike Professor Procrastinate, however, the emphasis in JONES is less on the temporal sequence of actions and more on the interplay between simultaneous actions—namely, steering and accelerating. This aligns with what Douglas Portmore refers to as "Maximalism", a position he associates with the PMH (Portmore 2018).

It is evident that there is a significant structural similarity between Troublemakers and the cases central to the actualism–possibilism debate, as well as between the challenges they pose. While both involve closely related issues, they differ notably in that the agents in the latter are temporal instances of the same individual across time. This structural parallel suggests that adopting a collective amendment in the context of Troublemakers implicitly implies, for the sake of consistency, a corresponding stance on the actualism–possibilism debate. Interestingly, in my search for comparable solutions, I discovered that Jackson himself proposed a resolution similar to Mixed Strategies and, by extension, to MAC (Jackson 2014).

Nevertheless, further exploration seems warranted due to the unique dependency between the temporal parts of agents, rooted in their shared (diachronic) personal identity—or whatever is the closest analog to that relation. This distinctive form of dependency could have meaningful implications, particularly for the relevant probability distribution (if one were to follow my recommendation to adopt MA-MOCOR), but potentially for other decision parameters as well. Examining these implications, especially in relation to the relaxation of the independence condition central to the analysis of collective decision situations, could lead to significant extensions of the approaches developed in this project. This prospect highlights a promising avenue for future research, offering both new insights and broader applications for the framework established here.

9.1.3 Generalization

I drove home many of my points using the simplest possible examples, focusing primarily on Coordination Cases and deliberately restricting myself to cases involving two agents with two options each. Additionally, I excluded Cumulative Effects Cases (a choice I believe is perfectly justified) and did not consider every conceivable collective amendment—perhaps, unintentionally, not even all the plausible ones. Nonetheless, this project invites further generalization along at least three dimensions.

Explicit Generalization Through Inductive Reasoning The cases I have discussed can be considered *base cases* in the sense of mathematical induction. This applies to scenarios like Two Factories, Seaman Clumsy, and Job Market. However, as the attentive reader will have noticed, my definitions were intentionally designed to be more general than strictly necessary. For instance, my formulations allow for the possibility that the consequences of individual actions in collective decision situations may themselves be collective decision situations. In the simplest cases, of course, the collective decision situation reduced by an action becomes an *individual* decision situation, which serves as a 'termination state', representing a final outcome.

In this sense, the *induction step* is already formally embedded in the framework. Executing it would be a straightforward, albeit laborious, task—one that would likely consume several dozen pages (or significantly more). I leave this exercise to future term papers or to those seeking to collect diligence points. Alternatively, it is an opportunity for anyone identifying errors in my claims to use them as a basis for straightforward publications—assuming, of course, that anyone even notices this book.

As often, things become interesting, if one allows *infinities* (cf. Hedden 2020). Both cases—the one involving infinitely many agents or actions and the other featuring agents with infinitely many options—open non-trivial avenues for exploration. For the former scenario, especially in cases of sequential action, discounting approaches may provide a workable solution. However, dealing with other forms of infinity in a systematic and orderly way poses a more complex challenge. Personally, I find the prospect of generalizing in this direction far more stimulating than the relatively mundane task of executing induction proofs.

Cumulative Effects Cases In Section 4.2, I justified excluding Cumulative Effects Cases from the scope of this project. The core of this decision lies in the substantial doubt surrounding whether these cases can be coherently or consistently described at all (cf. Kagan 2011). Simplifying somewhat polemically: how is the sum of many harms more harmful than … the sum of harms?

Adherents of MOAC may well ask this question with raised eyebrows. Nevertheless, it can be argued that at least some axiological theories (both narrowly and broadly construed, cf. Subsection 2.3.3) provide plausible accounts of such cases. Different arguments in this direction vary in persuasiveness (cf. Hedden 2020; Nefsky 2011; Spiekermann 2014). For instance, even if we can offer a convincing explanation of how the moral quality of outcomes can differ while remaining on par—say, within a hedonic, perceptual framework (cf. E. N. Dzhafarov and D. D. Dzhafarov 2010a,b)—consequentialists may counter by introducing novel, fitting aggregative amendments or by addressing probabilistic, vague trigger cases via probabilistic harms (cf. Shrader-Frechette 1987).

I do not dismiss the possibility of addressing these questions in the future. In fact, I have some preliminary work on this topic tucked away in my giant 'Archive of Unused Pages/Notes'. However, I am waiting for the right moment to bring it to light—perhaps when a suitable critique of MOAC emerges. Challenge MOAC effectively, and I will deliver the counter. After all, one need not fire all bolts at once. For now, this work remains safely stored in the metaphorical cellar of camp MOAC.

More Sophisticated Amendments Last but not least, one could also investigate further amendments. In my eyes, the most interesting ones would be those that try to include the fact that, often enough, not all agents act rightly. Maybe for an objective criterion of rightness, it is also perfectly okay to mark the reasonably best paths through our world full of imponderables. But perhaps, in the face of the newly accepted uncertainties stemming from Methodological Indeterminism, we should also make room for caution in the objective setting. Perhaps a *truly* reasonable formalization of PMH should accept a higher loss in the theoretically best case that, however, seldomly manifests in practice, if, in turn, more likely cases were better resolved. We would then abandon another idealizing assumption and, I guess, should again turn to decision theory to see what it offers. I think it would be particularly promising to try out *regret-based* amendments (cf. Bell 1982; Loomes and Sugden 1982) and, more generally, to make the notion of *risk* fruitful for objective consequentialist theories. Furthermore, I'd love to consider the details of an *ensemble solution* that allows one to choose between several amendments (or decision rules) on the features of the collective decision situation at hand (Chorus, Rose, and Hensher 2013).

More Sophisticated Amendments Last but not least, one could explore further amendments to enhance my multi-agent framework for consequentialism. Among the most intriguing, in my view, are those that take into account the reality that not all agents consistently act rightly. Perhaps an

objective criterion of rightness could, and even should, focus on marking the reasonably best paths through a world fraught with uncertainties and imponderables. However, given the uncertainties newly acknowledged through Methodological Indeterminism, it might also be worthwhile to introduce a greater degree of caution into the objective setting.

A *truly* reasonable formalization of PMH might, for instance, accept a higher theoretical loss in ideal cases that rarely occur in practice, in exchange for better resolving more probable, real-world scenarios. This would involve abandoning another idealizing assumption and, likely, turning once more to decision theory for guidance. One particularly promising avenue would be to experiment with *regret-based* amendments (cf. Bell 1982; Loomes and Sugden 1982), which emphasize minimizing regret over uncertain outcomes. More broadly, incorporating the concept of *risk* into objective consequentialist theories could yield valuable insights and refinements.

Additionally, I am especially interested in exploring the potential of an *ensemble solution*. Such an approach would involve selecting among multiple amendments (or decision rules) based on the specific features of the collective decision situation in question cf. Chorus, Rose, and Hensher 2013).

9.1.4 Formal Proofs

I claim that every major point I have made in this book is supported by argument. Of course, not every argument is developed in detail, and each could itself be supported by further reasoning. Inevitably, this leads to a regress that—while hopefully not infinite—-is characteristic of philosophical work, particularly within the analytic tradition. This is not a weakness unique to my work but a hallmark of the discipline.

There might, however, be another way forward. Twice during this project, my work almost turned into an extension of various stit-frameworks ("seeing to it that"; Belnap, Perloff, and Xu 2001; Chellas 1992; Horty 2001; Horty and Belnap 1995). Using a semantic framework for multimodal logics based on branching time models in the spirit of Prior (1955; 1967), I could have rigorously proven my arguments. These would involve an act operator (or, more traditionally, a "seeing to it that" operator), alongside temporal, modal, and deontic operators. Building on recent work by Horty and others (Horty 2019; Ramírez Abarca and Broersen 2021), I might even have introduced an epistemic operator. Over the span of two years, I produced nearly 250 pages of this kind of formalization, yielding several fruitful and, I think, informative proofs.

This material will undoubtedly see publication one day. Its relevance and appeal, however, depend on whether it can substantively reinforce the points made here, particularly those articulated in natural language and theorized

without formal (or, at least, axiomatic) systems. After all, as long as one progresses only within a formal system, one proves propositions solely within that system. For such work to become philosophically substantive, one would have to investigate to what extent findings from the formal system could be transferred to reality.

Because my primary aim was to say something substantial about the world—not merely about and within stit-semantics—I chose *not* to complete that formalization project. Instead, I returned to the approach presented here. Nonetheless, I believe that, starting from first principles and statements elevated to the status of axioms for describing relevant parts of reality (particularly in the context of consequentialist theorizing), it is possible to establish a sufficiently robust connection to formal stit-systems. Such a project, in my view, is ultimately worth pursuing and definitely belongs in the corpus of future work.

9.2 How the Tables Have Turned

The Challenge has long been considered *particularly* delicate for act-consequentialism for two primary reasons: first, because the commonly diagnosed inability to assess actions that, in combination, lead to morally suboptimal consequences strikes at the core of consequentialist intuitions; and, second, to borrow Kagan's phrase, because "consequentialism appears to fail even in its own favored terrain, where we are concerned with consequences and nothing but consequences".

I argue that my results fundamentally shift this narrative. The effective handling of collective decision situations should no longer be seen as a weak spot for consequentialism; rather, it emerges as one of its significant strengths. Consequentialists now possess a straightforward response to the challenges posed by the complex interplay of multiple interdependent agents.

This shift also has implications for 'the battle of the families of moral theories'. After all, adherents of deontological ethics also struggle with multi-agent scenarios, particularly in cases involving overdetermination. As David Killoren and Bekka Williams (Apr. 2013, p. 297) observe:

> Moreover, the *scope* of overdetermination problems seems to remain underappreciated. It is often assumed that overdetermination is mainly a problem for act-utilitarians (and other maximizing consequentialists), and that we are able to avoid such problems simply by shifting to a non-consequentialist view. That assumption is mistaken.

I couldn't put it better myself. Here is the example case they use to support their claim (ibid., p. 297):

Case 9.3 (Stooges) *Moe and Larry have promised to carry a piano upstairs by noon. This is a two-person job (neither stooge can carry the piano alone), and will require a half hour of time. Suppose that fulfilling this promise would result in greater utility than violating it [...] However, it is now 11:29 am, and both stooges are simply relaxing on the couch, each too lazy to put forth any effort.*

Here is, one last time, a normal form representing this case:

		Larry	
		chill	help
Moe	chill	promise goes unsatisfied, but both can chill (−)	promise goes unsatisfied, and Larry strains to carry the piano (−−)
	help	promise goes unsatisfied, and Moe strains to carry the piano (−−)	promise is fulfilled (+)

MAC can handle this case well: it is just another Troublemaker and we have derived that MA-MOCOR tells Moe and Larry that it is right for them to hold their promise. Thus, act-consequentialism performs better(now) than Killoren and Williams suggest.

Even if we conceive of this as a *true* overdetermination problem—assuming it is fixed that they will not fulfill their promise—the consequentialist faces no difficulty in resolving the situation appropriately. The key is to disambiguate what Killoren and Williams ask us to assume: "They have decided to let noon pass without even touching the piano". This scenario can be categorized into three distinct cases:

Collusion: If Moe and Larry colluded to avoid fulfilling their promise, then this collusion itself constitutes an upstream action that was wrong.

Sequential Decisions: If one of them decided first to relax on the couch, then the first agent to make this decision acted wrongly. A glance at the last GEF of the case (see Figure 9.2) illustrates this point.

Simultaneous Decisions: If both independently decided simultaneously to relax on the couch, then both acted wrongly. This scenario is also captured in Figure 9.2.

I think Killoren and Williams diagnose the challenge correctly when they write (ibid., p. 297):

> [T]he breaking of the stooges' promise is ensured by the fact that Moe refuses to keep it, and is ensured by the fact that Larry refuses to keep it. Thus, neither stooge is individually able to bring it about that their promise is kept. [...]

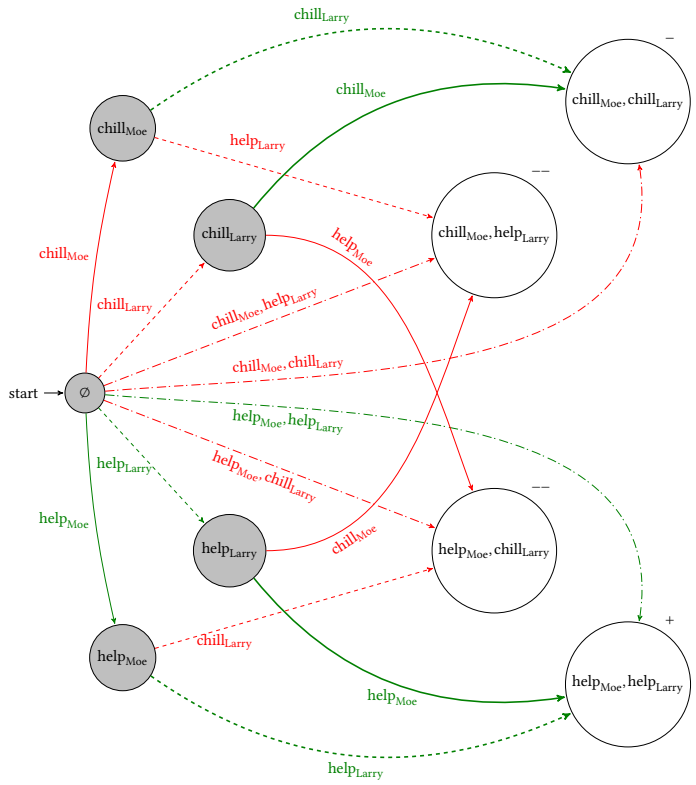


Figure 9.2: The GEF of STOOGES with annotations according to MAC. We can observe how the combination of both, Larry and Moe, helping move the piano is the right thing for them to do in the initial state.

> Neither Moe nor Larry is able to keep their promise, given that both Moe and Larry are unwilling to do so. Thus, it is unclear whether we can justifiably say that either of them is (as an individual) morally obligated to do so.

What, then, are non-consequentialists, particularly deontologists, likely to propose? Killoren and Williams advocate for an approach similar to Jackson's (1987; see also Subsection 4.4.3), though they avoid embracing Jackson's Difference Principle (i.e., Principle 3.1 in Section 3.3, page 49). This divergence allows them to distinguish between the obligations of individuals and the corresponding obligations of a constructed (or inferred) group agent. However, as we've already observed, such an approach generally challenges Methodological Individualism—a position that is far from comfortable.[164] This suggests that the CHALLENGE remains unresolved and underappreciated within deontological frameworks. In contrast, consequentialism holds a

[164] Killoren and Williams acknowledge this challenge but are willing to accept it (Killoren and Bekka Williams 2013, p. 304): "But we are willing to grant that [Moe, Larry] fails to exhibit such behavior, given its rampant irrationality. [...] Despite all this, we want to call [Moe, Larry] a moral agent that has various obligations."

clear advantage in addressing multi-agent scenarios—which, if we are honest, are ubiquitous.

Ultimately, I contend that the most promising path for deontologists lies in the *consequentialization* of their theories (cf. Dreier 2011; Hurley 2020; Portmore 2009). Such a move would allow them to utilize the APPROACH effectively, leveraging the newly identified consequences to infer rights, obligations, or other deontic constructs as they deem appropriate. Provocatively, I suggest that this would be a more honest strategy than inventing group agents where none exist or postulating *ad hoc* duties to cooperate in scenarios where cooperation was explicitly ruled out.

At camp Consequentialism, we welcome all comers. The more agents, the better—even if they are moral philosophers! After all, we are no longer afraid of multi-agent complexity; we've come to embrace it.

Bibliography

Andreou, Chrisoula (2014). "The Good, the Bad, and the Trivial." *Philosophical Studies* 169.2, 209–225.

— (2019). "Can Every Option Be Rationally Impermissible?" *Erkenntnis*.

Andrić, Vuko (2013). "Objective Consequentialism and the Licensing Dilemma." *Philosophical Studies* 162.3, 547–566.

Bacharach, Michael (1999). "Interactive Team Reasoning: A Contribution to the Theory of Co-Operation." *Research in Economics* 53.2, 117–147.

— (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton University Press.

Baier, Kurt (1958). *The Moral Point of View: A Rational Basis of Ethics*. Cornell University Press.

Bell, David E. (1982). "Regret in Decision Making under Uncertainty." *Operations Research* 30.5, 961–981.

Belnap, Nuel D., Michael Perloff, and Ming Xu (2001). *Facing the Future: Agents and Choices in our Indeterminist World*. Oxford University Press.

Bentham, Jeremy (2007). *An Introduction to the Principles of Morals and Legislation*. Originally published in London by T. Payne and Son, 1780. New York: Dover Publications.

Berkeley, George (1929). "Passive Obedience." In *Berkeley: Selections*. Ed. by Mary W. Calkins. Originally published in 1712. New York: Scribner's, 427–469.

Blyth, Colin R. (1972). "On Simpson's Paradox and the Sure-Thing Principle." *Journal of the American Statistical Association* 67.338, 364–366.

Bolander, Thomas (2024). "Self-Reference and Paradox." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2024. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2024/entries/self-reference/.

Braham, Matthew and Martin van Hees (2011). "Responsibility Voids." *The Philosophical Quarterly* 61.242, 6–15.

Brandt, Richard B. (1959). *Ethical Theory*. Prentice-Hall.

Briggs, R. A. (2023). "Normative Theories of Rational Choice: Expected Utility." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford Uni-

versity. URL: https://plato.stanford.edu/archives/win2023/entries/rationality-normative-utility/.

Broad, C. D. (1916). "On the Function of False Hypotheses in Ethics." *The International Journal of Ethics* 26.3, 377–397.

Broome, John (2021). "How Much Harm Does Each of Us Do?" In. *Philosophy and Climate Change*. Ed. by Mark Budolfson, Tristram McPherson, and David Plunkett. Oxford University Press, 281–292.

Brown, Campbell (2011). "Consequentialize This." *Ethics* 121.4, 749–771.

Budolfson, Mark (2019). "The Inefficacy Objection to Consequentialism and the Problem with the Expected Consequences Response." *Philosophical Studies* 176.7, 1711–1724.

Budolfson, Mark, Tristram McPherson, and David Plunkett, eds. (2021). *Philosophy and Climate Change*. Oxford University Press.

Bykvist, Krister (2002). "Alternative Actions and the Spirit of Consequentialism." *Philosophical Studies* 107.1, 45–68.

— (2003). "Normative Supervenience and Consequentialism." *Utilitas* 15.1, 27–49.

Castañeda, Hector-Neri (1968). "A Problem for Utilitarianism." *Analysis* 28.4, 141–142.

— (1974). *The Structure of Morality*. Springfield, Ill., Charles C Thomas.

Chellas, Brian F. (1992). "Time and Modality in the Logic of Agency." *Studia Logica* 51.3/4, 485–517.

Cholbi, Michael (2009). "The Murderer at the Door: What Kant Should Have Said." *Philosophy and Phenomenological Research* 79.1, 17–46.

Chorus, Caspar G., John M. Rose, and David A. Hensher (2013). "Regret Minimization or Utility Maximization: It Depends on the Attribute." *Environment and Planning B: Planning and Design* 40.1, 154–169.

Cohen, Daniel H. (1987). "The Problem of Counterpossibles." *Notre Dame Journal of Formal Logic* 29.1, 91–101.

Dietz, Alexander (2020). "Are My Temporal Parts Agents?" *Philosophy and Phenomenological Research* 100.2, 362–379.

Dong, Xibin et al. (2020). "A Survey on Ensemble Learning." *Frontiers of Computer Science* 14, 241–258.

Donne, John (1923). *Devotions upon Emergent Occasions*. Ed. by John Sparrow. Originally published in 1624; often referred to as Donne's Devotions. Cambridge University Press.

Dreier, James (1993). "Structures of Normative Theories." *The Monist* 76.1, 22–40.

— (2011). "In Defense of Consequentializing." In *Oxford Studies in Normative Ethics, Volume 1*. Ed. by Mark Timmons. Oxford University Press, 97–119.

Dzhafarov, Ehtibar N. and Damir D. Dzhafarov (2010a). "Sorites Without Vagueness I: Classificatory Sorites." *Theoria* 76.1, 4–24.

— (2010b). "Sorites Without Vagueness II: Comparative Sorites." *Theoria* 76.1, 25–53.
Estlund, David (2017). "Prime Justice." In *Political Utopias*. Ed. by Michael Weber and Kevin Vallier. Oxford University Press, 35–56.
Fehige, Christoph (1995). "Das große Unglück der kleineren Zahl." In *Zum moralischen Denken*. Ed. by Christoph Fehige and Georg Meggle. Vol. 2. Suhrkamp, 139–175.
Feldman, Fred (1980). "The Principle of Moral Harmony." *The Journal of Philosophy* 77.3, 166–179.
Ferreira, Jorge Viterbo (2018). "The Problem of Counterpossibles." MA thesis.
Gardiner, Stephen M. (2011). *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. Oxford University Press.
Gibbard, Allan F. (1965). "Rule-Utilitarianism: Merely an Illusory Alternative?" *Australasian Journal of Philosophy* 43.2, 211–220.
Glover, Jonathan and M. Scott-Taggart (1975). "'It Makes No Difference Whether or Not I Do It'." *Aristotelian Society Supplementary Volume* 49.1, 171–210. (Visited on 09/10/2020).
Gold, Natalie and Andrew M. Colman (2020). "Team Reasoning and the Rational Choice of Payoff-Dominant Outcomes in Games." *Topoi* 39.2, 305–316.
Goldman, Holly S. (1978). "Doing the Best One Can." In *Values and Morals*. Ed. by Alvin Goldman and Jaegwon Kim. Reidel, 185–214.
Gori, Marco, Alessandro Betti, and Stefano Melacci (2023). *Machine Learning: A Constraint-Based Approach*. Elsevier.
Guo, Xianping et al. (2009). *Continuous-time Markov decision processes*. Springer.
Gustafsson, Johan E. (2021). "Utilitarianism without Moral Aggregation." *Canadian Journal of Philosophy* 51.4, 256–269.
Gustavsson, Kent (2021). "Charlie Dunbar Broad." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2021/entries/broad/.
Harrod, R. F. (1936). "Utilitarianism Revised." *Mind* 45.178, 137–156.
Heath, Joseph (2020). "Methodological Individualism." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2020. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2020/entries/methodological-individualism/.
Heathwood, Chris (2020). "An Opinionated Guide to 'What Makes Someone's Life Go Best'." In *Derek Parfit's* Reasons and Persons*: An Introduction and Critical Inquiry*. Ed. by Andrea Sauchelli. Routledge, 94–113.
Hedden, Brian (2020). "Consequentialism and Collective Action." *Ethics* 130.4, 530–554.

Hooker, Brad (2023). "Rule Consequentialism." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Spring 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2023/entries/consequentialism-rule/.

Horty, John F. (2001). *Agency and Deontic Logic*. Oxford University Press.

— (2019). "Epistemic Oughts in Stit Semantics." *Ergo* 6.4, 71–120.

Horty, John F. and Nuel D. Belnap (1995). "The Deliberative Stit: A Study of Action, Omission, Ability, and Obligation." *Journal of Philosophical Logic* 24.6, 583–644.

Hurley, Paul (2020). "Consequentializing." In *The Oxford Handbook of Consequentialism*. Ed. by Douglas W. Portmore. Oxford University Press, 25–44.

Jackson, Frank (1987). "Group Morality." In *Metaphysics and Morality: Essays in Honour of J. J. C. Smart*. Ed. by J. J. C. Smart et al. B. Blackwell, 1–5.

— (1991). "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101.3, 461–482.

— (1997). "Which Effects?" In *Reading Parfit*. Ed. by J. Dancy. Blackwell, 42–53.

— (2014). "Procrastinate Revisited." *Pacific Philosophical Quarterly* 95.4, 634–647.

Jackson, Frank and Robert Pargetter (1986). "Oughts, Options, and Actualism." *The Philosophical Review* 95.2, 233.

Jeffrey, Richard (1982). "The Sure Thing Principle." In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. 1982. 2, 719–730.

Kagan, Shelly (2011). "Do I Make a Difference?" *Philosophy & Public Affairs* 39.2, 105–141.

Kavka, Gregory S. (1983). "The Toxin Puzzle." *Analysis* 43.1, 33–36.

Kemeny, J. G. and J. L. Snell (1960). *Finite Markov Chains*. Van Nostrand.

Killoren, David and Bekka Williams (2013). "Group Agency and Overdetermination." *Ethical Theory and Moral Practice* 16.2, 295–307.

Lawford-Smith, Holly and William Tuckwell (2020). "Act Consequentialism and the No-Difference Challenge." In *The Oxford Handbook of Consequentialism*. Ed. by Douglas W. Portmore. Oxford University Press, 634–654.

Lewis, C. I. (1932). "Alternative Systems of Logic." *The Monist* 42.4, 481–507.

Lewis, David (1973). *Counterfactuals*. Blackwell.

List, Christian and Philip Pettit (2011). *Group Agency: The Possibility, Design, Status of Corporate Agents*. Oxford: Oxford University Press.

Loomes, Graham and Robert Sugden (1982). "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty." *The Economic Journal* 92.368, 805–824.

MacFarlane, John (2023). "Belief: What Is It Good for?" *Erkenntnis*. (forthcoming).
Mackie, John Leslie (1977). *Ethics: Inventing Right and Wrong*. Penguin Books.
McCall, Storrs (1973). "Review: Selected Works by Jan Łukasiewicz, L. Borkowski." *Synthese* 26.1, 165–171.
McNamara, Paul and Frederik Van De Putte (2022). "Deontic Logic." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2022. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2022/entries/logic-deontic/.
Moore, George Edward (1903). *Principia Ethica*. Ed. by Thomas Baldwin. Mineola, N.Y.: Dover Publications.
Nash, John (1950). "The Bargaining Problem." *Econometrica* 18.2, 155–162.
— (1951). "Non-Cooperative Games." *Annals of Mathematics* 54.2, 286–295.
Nefsky, Julia (2011). "Consequentialism and the Problem of Collective Harm: A Reply to Kagan." *Philosophy and Public Affairs* 39.4, 364–395.
Notz, Dirk and Julienne Stroeve (2016). "Observed Arctic Sea-ice Loss Directly Follows Anthropogenic CO_2 Emission." *Science* 354.6313, 747–750.
Ord, Toby (2005). "Consequentialism and Decision Procedures." PhD thesis. University of Oxford.
Parfit, Derek (1984). *Reasons and Persons*. Oxford University Press.
— (1988). "What We Together Do." (unpublished), 1–33.
Pearl, Judea (2000). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.
— (2016). "The Sure-Thing Principle." *Journal of Causal Inference* 4.1, 81–86.
Peirce, Charles Sanders (1931). *Collected Papers*. Cambridge, MA: Harvard University Press.
Petersson, Björn (2017). "Team Reasoning and Collective Intentionality." *Review of Philosophy and Psychology* 8.2, 199–218.
Pettit, Philip and David Schweikard (2006). "Joint Actions and Group Agents." *Philosophy of the Social Sciences* 36.1, 18–39.
Pinkert, Felix (2015). "What If I Cannot Make a Difference (and Know It)?" *Ethics* 125.4, 971–998.
Portmore, Douglas W. (2007). "Consequentializing Moral Theories." *Pacific Philosophical Quarterly* 88.1, 39–73.
— (2009). "Consequentializing." *Philosophy Compass* 4.2, 329–347.
— (2018). "Maximalism and Moral Harmony." *Philosophy and Phenomenological Research* 96.2, 318–341.
— ed. (2020). *The Oxford Handbook of Consequentialism*. Oxford University Press.
Prior, Arthur N. (1955). *Time and Modality*. Greenwood Press.
— (1967). *Past, Present and Future*. Clarendon Press.

Prior, Arthur N. and D. D. Raphael (1956). "The Consequences of Actions." *Aristotelian Society Supplementary Volume* 30.1, 91–119.

Quinn, Warren S. (1990). "The Puzzle of the Self-Torturer." *Philosophical Studies* 59.1, 79–90.

Railton, Peter (1984). "Alienation, Consequentialism, and the Demands of Morality." *Philosophy & Public Affairs* 13.2, 134–171.

Ramírez Abarca, Aldo Iván and Jan Broersen (2021). "Stit Semantics for Epistemic Notions Based on Information Disclosure in Interactive Settings." *Journal of Logical and Algebraic Methods in Programming* 123, 100708.

Rawls, John (1971). *A Theory of Justice*. Belknap Press of Harvard University Press.

Regan, Donald H. (1980). *Utilitarianism and Co-operation*. Oxford University Press.

Savage, Leonard J. (1954). "The Foundations of Statistics."

Schelling, T. C. (1980). *The Strategy of Conflict: With a New Preface by the Author*. Harvard University Press.

Schnieder, Benjamin (2011). "A Logic for 'Because'." *Review of Symbolic Logic* 4.3, 445–465.

Schroeder, Mark (2021). "Value Theory." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2021/entries/value-theory/.

Schumpeter, J. A. (1908). *Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie*. Duncker & Humblot.

Shrader-Frechette, Kristin (1987). "Parfit and Mistakes in Moral Mathematics." *Ethics* 98.1, 50–60.

Simpson, E. H. (1951). "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 13.2, 238–241.

Singer, Peter (1980). "Utilitarianism and Vegetarianism." *Philosophy & Public Affairs* 9.4, 325–337.

Sinnott-Armstrong, Walter (2005). "It's Not My Fault: Global Warming and Individual Moral Obligations." In *Perspectives on Climate Change*. Ed. by Walter Sinnott-Armstrong and Richard Howarth. Elsevier, 221–253.

— (2022). "Consequentialism." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2022. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2022/entries/consequentialism/.

Sinnott-Armstrong, Walter and Richard B. Howarth, eds. (2005). *Perspectives on Climate Change: Science, Economics, Politics, Ethics*. Advances in the Economics of Environmental resources 5. Amsterdam.

Slote, Michael (1984). "Morality and Self–Other Asymmetry." *The Journal of Philosophy* 81.4, 179–192.
Slote, Michael and Philip Pettit (1984). "Satisficing Consequentialism." *Aristotelian Society Supplementary Volume* 58.1, 139–176.
Smart, J. J. C. (1973). "An Outline of a System of Utilitarian Ethics." In *Utilitarianism: For and Against*. Ed. by J. J. C. Smart and Bernard Williams. Cambridge University Press.
Smart, J. J. C. and Bernard Williams, eds. (1973). *Utilitarianism: For and Against*. Cambridge University Press.
Spiekermann, Kai (2014). "Causing Harm with Others: Small Impacts and Imperceptible Effects." *Midwest Studies in Philosophy* 38.1, 75–90.
Sverdlik, Steven (2011). *Motive and Rightness*. Oxford, UK: Oxford University Press.
Textor, Mark (2021). "States of Affairs." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2021/entries/states-of-affairs/.
Timmerman, Travis and Yishai Cohen (2020). "Actualism and Possibilism in Ethics." In *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2020/entries/actualism-possibilism-ethics/.
Timmons, Mark (2001). *Moral Theory: An Introduction*. Lanham, Md.: Rowman & Littlefield Publishers.
Toulmin, Stephen E. (1953). *An Examination of the Place of Reason in Ethics*. Cambridge University Press.
von Neumann, John and Oskar Morgenstern (1947). *Theory of Games and Economic Behavior*. Second Edition. Princeton University Press.
Weber, Max, Guenther Roth, and Claus Wittich (1978). *Economy and Society: An Outline of Interpretive Sociology*. University of California Press.
Wessels, Ulla (2002). *Die Gute Samariterin: Zur Struktur der Supererogation*. New York: De Gruyter.
Williams, Bernard (1973). "A Critique of Utilitarianism." In *Utilitarianism: For and Against*. Ed. by J. J. C. Smart and Bernard Williams. Cambridge University Press.
Williamson, Timothy (2002). *Knowledge and its Limits*. Oxford University Press.
Woodard, Christopher (2003). "Group-Based Reasons for Action." *Ethical Theory and Moral Practice* 6.2, 215–229.
— (2009). "What's Wrong with Possibilism?" *Analysis* 69.2, 219–226.
— (2019). *Taking Utilitarianism Seriously*. Oxford University Press.
Wroński, Leszek (2020). "Objective Consequentialism and the Plurality of Chances." *Synthese* 198.12, 12089–12105.

Zamir, Tzachi (2001). "One Consequence of Consequentialism: Morality and Overdetermination." *Erkenntnis* 55.2, 155–168.
Zimmerman, Michael J. (1996). *The Concept of Moral Obligation*. Cambridge University Press.
— (2014). *Ignorance and Moral Obligation*. Oxford University Press.

Declaration of AI Tools and Their Use

Three AI-supported tools were utilized in the preparation of this dissertation: DeepL for translation suggestions, ChatGPT (primarily the GPT-4-based version) for reformulation suggestions, and Grammarly as a Safari plugin for Overleaf to minimize grammatical errors. Importantly, ChatGPT was exclusively employed to help articulate my own thoughts, which I submitted as prompts. All AI-generated text was thoroughly reviewed, independently integrated into the manuscript, and adapted to meet specific context-dependent requirements. No long AI-generated text blocks were adopted verbatim.