TU WIEN Informatics

Choosing What to Believe

Belief Change Through the Lens of Rational Choice

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der Technischen Wissenschaften

eingereicht von

Adrian Haret, MSc
Matrikelnummer 01328338

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Stefan Woltran
Zweitbetreuung: O.Univ.Prof. Dipl.-Ing. Dr.techn. Thomas Eiter

Diese Dissertation haben begutachtet:

James Delgrande

Nicolas Maudet

Wien, 26. Juni 2020

Adrian Haret

TU WIEN Informatics

Choosing What to Believe

Belief Change Through the Lens of Rational Choice

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Adrian Haret, MSc
Registration Number 01328338

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Stefan Woltran
Second advisor: O.Univ.Prof. Dipl.-Ing. Dr.techn. Thomas Eiter

The dissertation has been reviewed by:

James Delgrande	Nicolas Maudet

Vienna, 26th June, 2020

Adrian Haret

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Adrian Haret, MSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 26. Juni 2020

Adrian Haret

Acknowledgements

Writing can be an excruciating affair, partly because one is forced to confront the gap between what is in one's head and what comes out on the page, and partly because LaTeXis so tedious. So the fact that this thesis managed to get done is proof that many stars were aligned, and many thanks are in order.

Most of the text to follow was written during January–April, 2020, and the influence of events unfolding across the world can be felt in the tone of the examples: from examples about Oscar nominees, conceived at the beginning of the year, to examples about doctors who have to decide on treatments for novel diseases, added in March. Conditions like these can be rattling at the best of times. In my case, they brought back memories of the Otto Wagner Spital (OWS) in Vienna, where I spent a few weeks in June, 2019. It is to the doctors and staff of the OWS that I want to extend my first thanks and gratitude.

The thesis itself grew out of successful collaborations and careful guidance. First acknowledgments go to my advisor, Stefan Woltran, who took a chance on me some years ago and hired me as a research assistant in his project, then guided me along the winding footpaths of academic research. Along the way there was help and encouragement: in the early days from Stefan Rümmele; in the latter years from Johannes Wallner, who so finely took up the role of employer, inspiration and colleague; and, not least, from Thomas Eiter, my second advisor.

Equally important were all the co-authors and collaborators I was fortunate to work with, listed here in chronological order: Thomas Linsbichler, Martin Diller, Jean-Guy Mailly, Jérôme Delobelle, Sébastien Konieczny, Julien Rossit, Andreas Pfandler, Nadia Creignou, Odile Papini and Martin Lackner. Thanks go out to the OeAD for funding my research stay at Univ. Paris-Dauphine, in Paris, during February–July, 2018, through a Marietta Blau grant. It was in Paris that I got my first taste of working on social choice topics, with Umberto Grandi, Arianna Novaro, Meltem Öztürk, Stefano Moretti and Hossein Khani. Special thanks go to Jérôme Lang, who hosted me in Paris; to Umberto, who hosted me in Toulouse; and to Nadia and Odile, who hosted me in Marseille.

Key in all this was the kindness and expertise of so many dedicated people, whose commitment to making things run smoothly was frankly invaluable: first and foremost are Juliane, Eva and Beatrix, along with the entire team behind the LogiCS program; then there is the team at the OeAD that coaches students in how to apply for their grants; and Eleni, at Dauphine.

Over the years I have been lucky to spend time with many splendid people, important for me in ways I cannot, and do not want to quantify, but still wish to acknowledge: Anna, Jan, Martin, Tobi, Thomas, Wolfgang and Zeynep, in Vienna, without whom lunches would have not been the same; my cousin Raluca, whose spare room in Vincennes I occupied in early 2018; Audrey, Emma and Harold, whose spare room in the 9^{th} arrondissement in Paris I occupied in mid-2018; Arianna and Dennis, in Toulouse; Hossein, at Dauphine.

Lastly, my gratitude goes out to Cătălin, a man for all seasons; my parents, who were there throughout it all; and Renate, who was there during the most difficult moments.

Kurzfassung

Ansätze der Wissensadaption die der AGM Schule folgen bilden eine wesentliche Grundlage um eine Vielzahl von Operationen, im speziellen in Bereichen in denen Agenten miteinander agieren, zu studieren. Bei der Untersuchung dieser Operatoren dient Aussagenlogik als lingua franca, nicht zuletzt um Bedingungen an diese Operatoren zu spezifieren, welche es erlauben rationale Operatoren von irrationalen zu unterscheiden. Aufgrund unterschiedlicher Auffassungen von Rationalität ergeben sich dadurch auch verschiedene und vernetzte Definitionen von Aspekten von Rationalität, die diese Verschiedenartigkeit widerspiegeln.

In dieser Disseration vertreten wir die These, dass Wissensadaption eine Verwandtschaft mit dem allgemeinen Problem der Entscheidungsfindung aufweist. Insbesonders muss ein Agent, oder eine Gruppe von Agenten, die Entscheidung treffen wie Wissen adaptiert werden sollte, wenn neue Informationen vorhanden sind und dabei sowohl die eigenen Positionen als auch die angesprochenen Bedingungen an Rationalität berücksichtigen. Beispielsweise sollte das Resultat der Adaption konsistent sein. Die Parallele zwischen Wissensadaption und Theorien zur Entscheidungsfindung wird durch den Gedankengang gestärkt, dass die jeweiligen Forschungsgebiete die gleichen Mechanismen für eine Auswahl von Optionen verwenden. Insbesonders werden gleiche Ansätze verwendet um zu zeigen, dass man die jeweiligen Operationen der Gebiete als rational ansehen kann.

Obwohl die Verbindung zwischen Wissensadaption und Entscheidungstheorie schon früher aufgezeigt wurde, argumentieren wir, dass noch viele Grundlagen offen sind. So können Operatoren der Wissensadaption als Entscheidungsprozesse angesehen werden, indem man die möglichen Resultate einer Adaption nach Präferenzen reiht. Durch diese Sichtweise können nicht nur Intuitionen die solchen Operatoren zugrunde liegen sichtbar gemacht werden, es können auch verschiedene Eigenschaften von verwandten Gebieten, die sich mit Theorien von rationalen Entscheidungen und der Sozialwahltheorie beschäftigen, auf Operationen der Wissensadaption angewendet werden. Ergebnisse solcher Untersuchungen führen, unter anderem, zu neuen Operationen in der Wissensadaption, welche auf diesen Eigenschaften beruhen.

Die wissenschaftlichen Beiträge dieser Dissertation sind zum einen eine Erweiterung der Studie von drei prominenten Familien von Operationen in der Wissensadaption: Revision, Update und Vereinigung von Wissen. Für alle drei Familien erweitern wir deren Anwendungsspektrum durch die oben genannten Ansätze. Für Wissensrevision schlagen

wir neue Postulate vor, die sich damit beschäftigen, wie bereits vorhandenes Wissen die Revision beinflusst. Für Wissensvereinigung, welche sich mit der Integration von Wissen im Kontext von Gruppen von Agenten beschäftigt, nützen wir Eigenschaften aus der Sozialwahltheorie, insbesondere solche Eigenschaften die sich damit befassen ob eine Operation manipuliert werden kann und solche die sich auf des Konzept der Proportionalität beziehen. Weiters erweitern wir vorangegangene Arbeiten im Bereich der Revision und des Updates von Wisssen welches in Hornlogik, einem Fragment der Aussagenlogik, formuliert ist. Wir sehen uns hierfür schwächere Varianten der gängigen Postulate an. Zum anderen beinhaltet die Disseration eine Untersuchung einer neuen Familie von Wissensadaptionoperationen, welche wir Enforcement nennen und als Grundlage für Revision von Präferenzen heranziehen.

In all unseren wissenschaftlichen Beiträgen verfolgen wir die in der Wissensadaption übliche Postulat-basierte Methodologie, um, im Einklang mit der oben genannten Verwandtschaft zur Entscheidungstheorie, diese Postulate als Wahlprozesse, sprich als Auswahlfunktionen von möglichen Resultaten, zu interpretieren. Durch Repräsentationsresultate zeigen wir, dass die Postulate durch Verwendung von Ordnungen rationalisiert werden können.

Abstract

Belief change, in the AGM tradition, gathers under a common methodological umbrella an array of operations, covering both single-agent and multi-agent processes. These operations are linked by the use of propositional logic as a *lingua franca*, are related by a network of interconnected rationality constraints and are united by the idea that they all describe, in some way or another, the dynamics of beliefs and information.

In this thesis we want to see belief change, thus construed, as akin to making a decision: according to this perspective an agent, or group of agents, faced with new information must make a decision as to what part of the new information to adopt, in a manner that balances both the agents' own positions, as well as certain rationality constraints, e.g., consistency. The parallel between changing a belief and making a decision is encouraged by the observation that both areas use the same underlying mechanisms of choices and preferences to rationalize their operations, in the process employing strikingly similar rationality constraints.

Though we are not the first to make this observation, we argue that there is still space to explore its implications. Seeing belief change operators as choice procedures that rely on preferences over outcomes allows us to tap both a series of useful intuitions about what belief change operators do, and a set of properties, scattered throughout the rational and social choice literature, that can aid the design of new instruments for belief change.

Thus, one side to our contribution to this thesis revolves around three existing prominent belief change operations, i.e., revision, update and merging, where these insights are employed in order to expand their range of application. For the case of revision we propose new postulates that deal with the way in which the prior information influences the revision process. For merging, which is a multi-agent operation, we adapt properties from the social choice literature, such as strategyproofness and proportionality, that formalize various aspects of fairness. We also look towards applications of revision and update to the Horn fragment of propositional logic, and extend existing work by studying weaker variants of the traditional postulates used in these cases. Another part of our contribution consists of a new type of belief change operation, which we call enforcement, and which we put forward both as a belief change operation in its own right and as theoretical background for a model of preference revision.

In all these cases we use the usual belief change tools of logical postulates, which we reinterpret, as part of our choice perspective, as constraints on a choice function over

possible outcomes. Through a set of representation results we are able to show, then, that the postulates can be rationalized in the familiar way using rankings on possible outcomes.

Contents

CHAPTER 1

Introduction

The theme of this work is that belief change, as it emerged from the seminal contributions of the late 1980s [Alchourrón and Makinson, 1985, Alchourrón et al., 1985, Gärdenfors, 1988], and rational choice, as shaped in the foundries of decision theory a few decades earlier [Neumann and Morgenstern, 1944, Arrow, 1951, Arrow, 1959, Sen, 1970, Suzumura, 1983], have much in common. This similarity, we argue, has two aspects to it.

At first blush, there is the simple observation that belief change and rational choice can be seen to share a common mathematical framework: indeed, a peek at the formal structure shows that, beyond the distinct motivation and the different terminology, the methodology of one area dovetails nicely with the methodology of the other. The standard recipe for approaching a belief change operation is to propose a set of appealing normative properties, usually in the form of *logical postulates*, and, in parallel, to look for constructive ways of satisfying these postulates; often this is aided by the illuminating instrument of a representation result, which serves as a bridge between the postulates and a broad range of constructive procedures put forward. On its side, rational choice proceeds along the same lines: there are abstract axioms describing reasonable choice procedures, concrete decision mechanisms, and characterization results showing how to relate the two. Both fields, in short, rely on the axiomatic method to understand the objects of study. What is more, some of the main concepts used also bear a striking semblance to each other: belief change usually involves agents who rank possible states of the world in terms of their plausibility: an agent changing its mind, we will see, is required to make a judgment over what is more likely to be the case given some information it receives. These rankings, then, can be equated with the preferences that decision makers, in rational choice, have over alternatives: formally, thinking that an outcome is more plausible than another and saying that an alternative is better than another are modeled by the same types of relations.

The existence of this common ground, we propose, opens up a wellspring of techniques that can be exploited to mutual benefit, and the present thesis is devoted to exploring the

consequences of this particular viewpoint. Though belief change is ostensibly not about choice, the formal similarity alone warrants the expectation that some traffic of ideas would be valuable. And, while the results offered here are focused mostly on how rational choice informs belief change, some glimpses of how the relationship can be reversed are scattered along the way.

Going further, we suggest that belief change and rational choice share not just a mathematical skeleton, but also a common intuition. That is to say, it is not just that both areas use postulates and representation results to structure their content: but the postulates express the same principles, and the representation results are used to characterize one single, underlying mechanism.

Rational choice

The mechanism is that of *choice*, and the principles surrounding it are the conditions that, in the theory of decision making, guarantee that choice can be thought of as rational. These are the same principles underlying the economic view of *homo economicus* as maximizer of utility, and that have been used as a starting point for much research in the foundations of economic theory.

Choice, here, covers both the individual case, i.e., that in which one agent has to choose among a set of alternatives, with the prototypical example being that of a decision maker choosing a bundle of goods subject to a budget constraint, as well as the social case, i.e., that in which a collective of agents has to choose among a set of alternatives, with the prototypical example being that of a society organizing an election.

Example 1.1: Treating a novel disease

A doctor, who we will call here Doctor 1, is on the frontlines of medical assistance for a newly discovered respiratory disease. The doctor has to choose among four possible ways to treat patients diagnosed with this disease. At the moment of the decision there is no clinically proven treatment for the disease, but a and b, two existing drugs, have shown a certain amount of promise. The doctor has to decide between the option of administering both drugs together, only one, or neither. After some deliberation, the doctor chooses to administer both a and b, only to be told that b is momentarily out of stock. The doctor goes ahead and administers a.

Later in the week, all five doctors in the infectious disease wing of the hospital meet to decide on a common strategy for how to treat the disease. Doctor 1 has the following to say:

Doctor 1: Based on my understanding of this disease, drug a has the highest chance to work against this disease, slightly higher than b even. But a and b probably have the largest impact when administered together. In any case, it seems to beyond doubt that administering either is preferable to doing nothing.

But, in general, opinions are all over the place:

Doctor 2: I am more concerned with the possible harm using untested drugs might have on our patients. Since there have been no clinical trials on these drugs, I think there's no good ground to push either of them, b more so than a; adding a might make the combo less potent, but I would think that the safest bet is to use neither.

Doctor 3: I'd say that drug b is a good bet, and using it is probably better overall than leaving it out of the treatment plan. But a is known to cause certain bad side effects if administered in the wrong dosage, and mixing it in would maybe harm the patient.

Doctor 4: Drug a has been shown to be effective for other diseases, so we might give it a shot here as well; giving b might be better than doing nothing, but I think a and b should definitely not be mixed.

Doctor 5: Drug a sounds good, and maybe in combination with b it might be effective, certainly more effective than doing nothing. But drug b is already in use for other diseases, and basing therapy just on it might create a shortage of b for those patients. So I am strongly opposed to using b by itself.

Settling on one course of action will be difficult.

Example 1.1 shows two cases where choices need to be made. One is the single-agent case of Doctor 1 choosing among four different treatment options, then having to adjust the choice to stay within the 'budget' of what medication is actually on stock. The second is a multi-agent case in which a choice about the treatment is made on the basis of input from several doctors: each doctor, presumably, brings with them experience and expertise, and their opinions are to be treated equally. In both cases a decision needs to be made: all these options, what to choose? And, in both cases, the challenge is to come up with a formal model of what it means for a decision to be *good*.

Most existing work in rational choice can be traced back to a model of what makes for a good decision that, in its basic form, says: always choose the best alternatives available. This is a view that, at first glance, does not seem particularly informative, but it carries some important implications. First of all: good for whom? If the decision concerns only one agent, then ostensibly that agent gets to have the final say: when Doctor 1 in Example 1.1 has to choose by themselves, then the choice depends only on their own assessment of what would work best. But if the decision concerns a group of agents, then the answer can vary. In Example 1.1, it is not obvious what the collectively best option is: perhaps the majority should prevail; maybe priority should be given, instead, to the worst-off agents; or perhaps the spoils should be distributed proportionally, in accordance with the composition of the group. There is no pretense that any of these is

the undisputed right answer: social choice is committed to studying all these options, and we will see that they are equally relevant for multi-agent belief change.

Second, what does it even mean for something to be good? To avoid the matter becoming too perplexing, a formal analysis might only consider an ordinal ranking of alternatives: how good alternatives are, not in absolute terms, but only relative to each other. This makes it possible to identify the best alternatives, without having to go into metaphysical details about what makes something good, or desirable. But, though easier to model, this assumption still requires an impressive feat: the agent (or collective of agents) must be able to pick out the best alternatives out of a given lineup, i.e., it (or they) must possess the ability to rank alternatives according to their quality, and to do so in a way that is coherent across the set of alternatives. This is where preferences come in: in Example 1.1, each doctor has their own assessment of how the various treatment options stack up against each other. These assessments differ in their motivations: Doctor 1 evaluates the treatments in terms of which has the highest chance of being effective, whereas Doctor 2 evaluates them in terms of their likelihood to cause harm, while the other doctors land somewhere in between and use a variety of other criteria. This use of different scales means that judging such preferences in absolute terms would be difficult, if not, as has been claimed elsewhere, meaningless [Arrow, 1951]. But through the intermediary of an ordinal ranking, the preferences can be put side by side and aggregated.

How can we rationalize that an agent prefers an alternative x_1 over an alternative x_2? Does it make sense to say that a society, as a whole, prefers x_1 over x_2? Such questions lie at the heart of rational agency and democratic decision making. Showing that an ordinal ranking of alternatives, which we will call here a *preference order*, exists and has desirable properties turns out to be a far from trivial matter, and significant contributions in rational choice have focused on the conditions under which it can be settled: this includes results in the theory of individual choice on the constraints that have to be imposed on choice behavior to guarantee conformance with a desirable preference order [Sen, 1969, Sen, 1970], or Arrow's celebrated impossibility theorem in social choice showing that under mild assumptions such a preference order does not even exist [Arrow, 1951].

The insight that emerges from these results, and that we will exploit in the present work, is that, in a minimally rational agent, there is a tight connection between preference and choice. This connection runs in two directions. On the one hand, preference can be taken to determine choice, by making clear what the best alternatives are: if an agent believes alternative x_1 is better than x_2, then, faced with a choice between the two, and all thinge being equal, the agent will choose x_1. On the other hand, preference can be inferred from choice, by taking choice behavior to be indicative of what the best alternatives are considered to be: if an agent chooses x_1 over x_2 when both alternatives are on the table, then that must be because the agent thinks x_1 is better than x_2. The key question is whether the two directions can be shown to cohere with one another: if the agent chooses x_1 over x_2 once, x_2 over x_3 later, and x_1 over x_3 at yet some other time, then all is well; this behavior is consistent with our inference that the agent thinks x_1 is better than x_2, x_2 is better than x_3, and hence x_1 is better than x_3. But if the agent chooses

x_3 over x_1, then something has gone wrong: this indicates that there is a cycle in the agent's preference, and our intuitions about what constitutes rational behavior have been violated.

In this work we want to focus mainly on what has to happen for things to go smoothly, i.e., the conditions under which there exists a preference order driving choice behavior, this order is transitive (or, at the very least, avoids cycles), and satisfies whatever other properties we find appealing. A more detailed exposition of these properties, together with the constraints on the choice function needed to make them happen, is offered in Section 2.4.

The mark of a rational agent, in this setting, is not so much in what, concretely, it prefers, but that, whatever the agent's preferences are, choices made on their basis are coherent and successful. Formally, we know we are on the right track when the setup created allows us to derive a *representation theorem*, i.e., a result saying that choice under whatever axioms we have come up with is equivalent to having an intended type of preference order over alternatives and selecting the best alternatives on offer. Ultimately, a representation theorem validates the view according to which rational choice is equivalent to optimizing over the space of alternatives, and in Section 2.4 we will get a glimpse of some significant representation results for choice functions.

But what does all this have to do with belief change?

Belief change

Belief change, on a first approximation, is concerned with epistemic states, i.e., the information that agents hold in their 'heads' at any given moment, and their dynamics, i.e., the ways in which epistemic states are supposed to change in light of new information. The main types of belief change we will focus on in this work are the established operations of revision [Katsuno and Mendelzon, 1992], update [Katsuno and Mendelzon, 1991], and merging [Konieczny and Pino Pérez, 2002, Konieczny and Pino Pérez, 2011], as well as the relatively new operation of enforcement [Haret et al., 2018c]. Revision, update and enforcement study the dynamics of a single agent's epistemic state, whereas merging studies the dynamics of the epistemic states of groups of agents. All these operations are instances of a formal framework that has been in place since the seminal first contributions to the field [Alchourrón et al., 1985, Alchourrón and Makinson, 1985, Gärdenfors and Makinson, 1988, Gärdenfors, 1988], but whose main ideas certainly go back even further [Harper, 1976, Levi, 1980]. The main tools of this framework are a language, usually propositional logic, in which beliefs are written down; a set of postulates that capture our intuitions about how a rational belief change operator should behave; and a set of concrete procedures that perform the operations, validated by representation results.

Nominally, all these operations study the dynamics of beliefs: it is customary to speak of *belief* revision, *belief* update, and so on, and many motivating examples are constructed around agents that find out they are wrong, or that the world has changed. However, as a note for what is to come, we believe it is important not to read too much into this

term: indeed, the barebones AGM model of revision usually taken as a reference point for work in belief change [Alchourrón et al., 1985], is flexible enough to account for the dynamics of not just beliefs, but also, possibly, intentions, goals, or desires: in general, any type of attitude towards an item of information that makes the agent partial, in some way, to that information. Though some parallels between the 'beliefs' of belief change and beliefs as studied in philosophy [Schwitzgebel, 2019] are welcome, and we will draw such parallels here too, we do not want to be particularly dogmatic about the meaning assigned to the inputs and outputs of belief change operators. And certainly, nothing in the AGM model, or its various offshoots, commits us to any particular metaphysical doctrine of what these inputs are. For the rest of this work the beliefs we will look at are a moving target, and the only uniform assumption we want to make about them is that they are particular to an agent, and the agent is interested in seeing them come to pass. In this respect, we want to say, change of beliefs becomes something like a decision problem studied in the field of rational choice.

Belief change and rational choice

It is by now fairly well documented that revision is guided by similar principles as those that govern rational individual choice [Doyle, 1991, Rott, 1992, Bonanno, 2009, Arló-Costa and Pedersen, 2010, Ma et al., 2015], with some of the prototypical representation theorems [Gärdenfors and Makinson, 1988, Grove, 1988, Katsuno and Mendelzon, 1992] showing that revision operators can be seen to rely on preference orders in just the same way that choice functions do. Likewise, it has never been a secret that merging has many points in common with social choice, with several existing works, including some that provide the material for this thesis, dedicated to exploiting the connection between the two [Eckert and Pigozzi, 2005, Everaere et al., 2007, Gabbay et al., 2007, Everaere et al., 2014, Lang and Xia, 2016, Haret et al., 2016b, Haret and Wallner, 2019, Haret et al., 2020].

The connection, we want to say upfront, is straightforward. Revising a belief is like choosing from the outcomes consistent with the new, accepted information. Which outcomes? The ones that are, of course, 'best' according to some preference order. Merging is aggregating information provided by different agents, similar to how individual preferences in an election would be aggregated. But, though this viewpoint can be summarized in a few lines, its implications, we think, are yet to be fully explored. Our aim in this work is to build on the connection between belief change and rational choice and use it to both make sense of some existing work in belief change—not just revision and merging, but also update and enforcement—and also to suggest new avenues of research. The guiding thought is that seeing belief change as a form of choice opens up an entire raft of new and exciting possibilities: from a rich set of properties that have traditionally been used to understand decision procedures, individual and collective, and that can be brought to bear on existing belief change operators, to an assortment of tools that can facilitate the application of belief change techniques to novel contexts. Making the choice component of belief change explicit brings with it clarity and ideas for how to

go further. Or so we will argue.

In our view, the choice perspective is relevant to the belief change operations mentioned above in three main ways. First, it provides a good set of intuitions for understanding aspects of belief change that have been present since the AGM beginnings, such as the role and interaction between the basic postulates: the motivation behind the more arcane postulates for revision, for instance, becomes more vivid when interpreted in terms of coherence of choice; the difference between revision and update becomes clearer when the operators are reinterpreted as choice functions over possible worlds; and the mechanics of a merging operator is illuminated by understanding that merging can be seen as an election whose aim is to decide the correct bits of information and be fair towards the participants. All these aspects are presented in more detail in Chapter 3.

Second, the choice perspective suggests new tools for analyzing familiar notions: the literature on rational choice is rich with distinctions and properties that can be readily adapted to the context of belief change, if we allow ourselves to see belief change operators as social procedures that rely on (some type of) preference rankings. The applications we have in mind include tweaking the way in which rankings are generated for revision to reflect the agent's bias towards its own belief, which we look at in Chapter 4, or coming up with new postulates for merging aimed at capturing various aspects of fairness, which we study in Chapter 5. Notably, the literature on rational choice also anticipates the acyclicity postulate needed to make Horn revision work [Delgrande and Peppas, 2015, Delgrande et al., 2018], in the form of *Suzumura consistency* [Suzumura, 1976, Suzumura, 1983, Bossert and Suzumura, 2010]. This will be discussed in Chapter 6.

Finally, seeing belief change operators as choice functions over the set of possible worlds provides guidelines for what to look for when exporting the formalism of belief change to other contexts, in this way paving the way for new applications of belief change. We will see this mindset at work in Chapter 6, where we try to understand revision and update applied to Horn formulas, and where the property of Suzumura consistency will serve as a guide to pinpointing the exact type of preference order we need to capture, and in Chapter 7, where we look at revision of preference orders.

Choosing what to believe

The line of reasoning we want to pursue in this work can be better grasped by looking at some examples of belief change in action, with no better place to start than the paradigmatic case of *revision*.

Example 1.2: A monopoly on tool use?

It used to be widely believed, among primatologists, that humans were the only animals to use tools. Then, in the 1960s, a young researcher studying chimpanzees in the Gombe Stream National Park in Tanzania, observed a male chimpanzee named David Graybeard, using straws and intentionally stripping branches to fish for

termites. Here was a non-human primate using tools in an undoubtedly sophisticated way, something considered impossible at the time. The researcher, who was none other than Jane Goodall, was on track to overturn a centuries old orthodoxy.

After some tussle with getting her work published and acknowledged by the primatology community, Jane Goodall's discoveries were finally accepted. But incorporating these discoveries into the corpus of existing data posed a dilemma, for what Jane Goodall had unearthed did not fit with the rest of the beliefs in place. In a telegram, prominent British paleonathropologist Louis Leakey wrote to Goodall: "Now we must redefine tool, redefine Man, or accept chimpanzees as humans." [Goodall, 2010]

Example 1.2 already announces some of the intricacies of belief change that a formal analysis will have to contend with. It highlights that beliefs, no matter how entrenched, may come under scrutiny by new discoveries; that, if such discoveries are accepted, they might interfere with older beliefs; and that, in order to maintain sanity, some of the old beliefs must be reshuffled to accommodate the new ones. The basic, no-frills model of revision we will present in Section 3.1 captures only the final step in Example 1.2: that in which the new information has already been vetted and accepted, and the existing beliefs need to make space for it.

Example 1.2 also anticipates, through Louis Leakey's telegram, the theme of this work: that belief change often involves some kind of choice, in this case over what items of the prior beliefs to give up. Indeed, Leakey seems to suggest that the best response to Jane Goodall's finding is either to redefine the concept of tool or that of human being, (presumably, to give up the notion that humans use tools), or to accept chimpanzees as humans (presumably, to give up the notion that chimpanzees and humans are different species). Either of these possibilities, we are led to understand, were thought by Leakey to be more plausible than simple acceptance of the fact that a non-human animal could be capable of tool use. Though this sounds like an odd revision policy, and it is likely that it was put forward as a joke, it illustrates how the process of reshuffling opens up several possibilities, and that resolving them requires an act of choice. We will elaborate this aspect in Section 3.1, where we will see that revision can be understood as choosing the most likely outcomes from the ones considered feasible. All this expository work will rely on existing, classical results, our only input being to draw attention to the way in which revision can be reconceptualized as a sort of decision the agent has to make about what to keep and what to give up among its most cherished states of the world. In doing so, we will also see that standard models of revision rely on a particular assumption, intended to cover cases when no choice needs to be made: namely, that if the new information does not contradict the prior beliefs, then revision amounts to simply incorporating the new information into the existing beliefs. This will allow us, in Chapter 4 to see that such an assumption involves an element of choice architecture (i.e., about how the prior beliefs are prioritized in the revision process), that it is not always warranted, and we will look at alternative ways of modeling it.

To be clear, the claim we are making is not that revision is about believing whatever one wants to believe, or that when doing revision an agent shops around for a new belief and just settles on the one that sounds nicest. Rather, our point is that the cognitive mechanism underlying revision is very similar to the mechanism underlying rational choice. At the basic level this is simply because, as we will see in Sections 2.4 and 3.1, the mathematics is very similar; but we want to go further than that and suggest that this is not a coincidence: if the posterior information is not determined purely from the prior and new information, as in Example 1.2, i.e., if there is more than one possibility for what the posterior information can be, then some sort of selection has to happen if any type of reasoning, or inference is to take place; and a natural way of describing this process is using the language of choice and preference: when the agent has to form a new belief it, or we may say, its mind, will select the information that best fits the new data it is dealing with, i.e., it will make a choice informed by a preference. In all this, it is important to keep in mind that we intend the notions of choice and preference to be tightly interconnected, i.e., given the preferences, then choices are completely determined, and likewise in the opposite direction. Indeed, a large part of our efforts will be dedicated to making sure that preferences and choices, in this sense, fit seamlessly with each other. And it is also important to keep in mind that we intend the notion of preference to cover a large amount of ground: in a cognitive setting, as befits revision, preferences can encode something like the agent's evaluations of how likely outcomes are; or, their salience in the agent's mind; or, the order in which the agent would like to see them occur. All these aspects describe *bona fide* attitudes an agent can have with respect to the outcomes, and it is part of the belief change project that their dynamics can be brought under the umbrella of one formal model.

That being said, an equally important observation is that there is not one single type of rational belief change: rationality comes in many flavors, and different situations call for different approaches. Thus, the final sentence of the previous paragraph would be more precise if it called for *a family* of formal models, rather than just one model. To some degree this insight was already present in the original model, with the distinction between contraction, expansion and revision [Alchourrón et al., 1985, Gärdenfors, 1988]. It was soon ovious that this distinction did not exhaust all the types of belief change worth studying, and subsequent research has seen a proliferation of belief change operations adapted to various use cases. In this work we will focus on a small sample of such operations.

Example 1.3: Keeping up with the humans

My home is controlled by a software assistant, the latest in AI smarthome technology. Most of the things it does are with a clear mandate from myself. I tell the assistant to keep the temperature above 15° C and, in a desire to unplug during the evenings, to turn the Wi-Fi off starting with 21:00. The assistant is receptive to my instructions, which usually come in the form of *if this, then that* statements, and implements them

most dilligently. But it gets even better: the assistant follows my every move, trying to guess my needs and wants, and adapts its actions to what it perceives are my behavior patterns.

Thus, after repeated exposure, the assistant learns that I sometimes turn the Wi-Fi back on after 21:00, and this usually coincides with times when my friend, who lives on a different continent, is online. After a while, the assistant asks me if it should integrate this information into its rule base and I agree. The assistant modifies its list of instructions accordingly, keeping the Wi-Fi on during the evenings when, and only when, my friend is online.

Example 1.3 features an artificial agent whose epistemic state consists of rules about how it is supposed to manage a household. The input for the agent comes from observations it makes, and we may assume it is as reliable as can be: for instance, the agent only accepts new facts after they have undergone a long enough process of confirmation. The result, then, is a change in the rules that the agent implements: at first glance, a case of revision not unlike the one described in Example 1.2. Note, however, that in this case the new information (i.e., that the Wi-Fi is on if and only if my friend is online) is not inconsistent with the rules already in place; and, if the assistant were to simply add the new information to its epistemic state, as we said revision is committed to doing, then it would use its prior information that the Wi-Fi should be turned off after 21:00 in conjunction to the new information to infer that my friend must actually be offline!

The scenario described in Example 1.3 shows that sometimes adherence to prior information at all costs is wanted. In these cases we would expect that the posterior information retains more of the new information than what would be warranted by a revision operator, while still being biased by whatever prior beliefs were in place. Such an operation is that of *update* [Katsuno and Mendelzon, 1991], and we will see in Section 3.2 that the principles behind it are are subtly different from the principles behind revision. We will also see that this difference can be elegantly expressed in the way that an update operator chooses what to retain from newly acquired information.

The assistant in Example 1.3 reasons in terms of observations and rules: *if this, then that*. We can assume this is because it is a fitting mode of thought for an AI assistant and, not least, because it keeps the complexity of reasoning within a manageable limit. In general, we can imagine that belief change, be it revision or update, is done by agents with limited expressive and computational resources, in languages that are specialized to a particular application. What this means, in practice, is that the epistemic states of the agent, i.e., its prior and posterior information, and possibly the incoming information as well, need to adhere to some specific format. And it is important, if belief change is to be applied outside its propositional logic ivory tower, that the main insights can be exported to other languages: ideally, these insights end up holding in the specialized formalisms in the same way that they hold in the base language of propositional logic. Experience shows, however, that the choice of language makes a significant difference to

what a belief change operator can do, with some effort having to be expended just so we can arrive at the same results.

In Chapter 6 we will look at revision for Horn formulas: such formulas make up a language that can be seen as a restricted fragment of propositional logic, and that is suited to represent facts and rules such as the ones my (hypothetical) AI assistant uses. Chapter 6 will start by surveying the measures that need to be taken to emulate the classical representation results for revision with Horn formulas. We will see that belief change for such a formalism works as a form of *constrained* choice, constrained by the strictures of the language we are working in. And we will also see that existing insights for how to work around these strictures have clear analogues in properties that have been studied in the rational choice literature; the same properties, then, allow us to extend these results to update for Horn formulas. This is an example of how sensitivity to the way in which choice guides belief change can suggest fixes when the usual techniques break down.

In both Examples 1.2 and 1.3 it is assumed that the newly acquired information comes from an authoritative source, and that it takes precedence over the prior information if a conflict between the two is present. But we can also imagine cases where the newly acquired information stems from a source that is trusted, though not necessarily more than the prior information.

Example 1.4: The art of diagnosis

A doctor sees a patient presenting with cough and a stuffy nose. Based on an initial examination, the doctor concludes that it is either an allergic reaction, or bronchitis. The patient, who has done their own research of the symptoms online, points out that the symptoms are consistent with a new strand of coronavirus that has been making the headlines. After checking this information the doctor acknowledges the possibility, and accepts that it can be one of the causes for the patient's symptoms. Thus, the doctor becomes committed to take the possibility of a coronavirus diagnosis possible, but does not consider it strictly more likely than their own original assessment. As a result, the doctor just adds the coronavirus hypothesis to the other two conditions consistent with the symptoms, i.e., allergies or bronchitis.

Example 1.4 calls for a type of change that, to the extent possible, treats newly acquired information as equally likely as the prior information. This is a different strategy than that employed by either revision or update, where new information is accepted even if this comes at the cost of giving up the prior information entirely: here we want to preserve as much of the prior information as possible, alongside the new information. In choice terms, this is equivalent to deciding not which outcomes consistent with the new information to remove, but which outcomes consistent with the prior information to add; ideally, as in Example 1.4, we can add all these outcomes and expand the epistemic state to one that gives equal weight to all of them. But if, in doing so, the answer grows to include all possible outcomes, and thus becomes non-informative, then a choice becomes mandatory and the preference mechanism kicks into gear. We call this new type

of operation *enforcement* [Haret et al., 2018c], and in Section 3.3 we will see that the principles behind it are yet more different from the principles behind revision and update. Enforcement provides an example where seeing belief change as a form of choice proves useful in understanding a novel type of operator.

Examples 1.2, 1.3 and 1.4 all track changes in the epistemic state of one agent: the prior beliefs and the posterior beliefs, for these operations, are always situated in one agent's 'head'. We can easily imagine, however, that several agents pool their information together in the attempt to reach a common conclusion, as is the case in a group decision scenario. There are many instances of this type of aggregation, from nationwide elections to the group of doctors in Example 1.1 that have to converge on a common treatment protocol. A more showy, if less serious, example is the decision process leading up to the annual list of Oscar nominees. Though we do not know the precise details of how this process works, we can consider its complexities in a toy example.

Example 1.5: #OscarsSoFossilized

Year after year, the Oscars attract the ire of moviegoers everywhere for their choice of who to acknowledge. Sometimes this is for handing out honors to people who, it is thought, do not deserve them; at other times it is for ignoring people who do; usually it is for both, and the year 2020 was no exception [Brody, 2020]. For simplicity, assume the Academy consists of exactly four members, who have to decide the 2020 nominees for the category of Best Director. There are three directors up for contention: Alma Har'el, director of *Honey Boy*, Bong Joon Ho, director of *Parasite*, and Céline Sciamma, director of *Portrait de la jeune fille en feu*. The final lineup is supposed to consist of exactly two nominees, so not all of the three names can make the cut. When queried, the Academy members express the following opinions:

Member 1: I think Alma Har'el and Bong Joon Ho should be nominated. I haven't seen *Portrait de la jeune fille en feu*, so I have no opinion on Céline Sciamma.

Member 2: Definitely Alma Har'el, and maybe Bong Joon Ho or Céline Sciamma too.

Member 3: I think only Bong Joon Ho deserves the nomination.

Member 4: Neither Alma Har'el nor Bong Joon Ho seems good enough, but Céline Sciamma's movie really impressed me and I think she should be nominated.

Since there can be only one list of nominees, the opinions of the four members need to be aggregated into a consensus opinion. This consensus has to reflect the opinions that go into obtaining it, and it has to meet the size constraint, i.e., that there can be only two nominees.

After some back and forth, during which they realize that there is no way of making everyone happy, the Academy members decide to nominate Bong Joon Ho and Céline Sciamma, on the grounds that this will lead to the least amount of unhappiness.

There is a split, in Example 1.5, between the four Academy members, i.e., there is no lineup universally agreed upon. A choice needs to be made, but whatever it is, someone will be unhappy. Can a fair outcome be ensured? Example 1.5 illustrates some of the key challenges of aggregating information originating from different sources: the aggregation procedure should balance each source in an appropriate way, e.g., by factoring in reliability of the sources if the goal is an accurate result, or, on the contrary, by treating all sources equally if, as is the case in Example 1.5, the goal is a fair result; the sources may provide conflicting information, such that there is no one answer that fits all; the information provided may reflect complex interdependencies between issues (if this, then that, and if not then perhaps some other thing), which adds an extra layer of complexity to the issue; the result may be expected to meet certain additional criteria, e.g., it should be of a specific format, or, as in Example 1.5, satisfy certain cardinality constraints. These challenges go beyond the challenges of deciding with just one agent, and new techniques need to be brought in.

We will model tasks like the one faced by the Academy members in Example 1.5 using the framework of *merging* [Konieczny and Pino Pérez, 2002, Konieczny and Pino Pérez, 2011]. In Section 3.4 we will present a set of established principles for thinking through such scenarios, together with a handful of mechanisms for extracting an actual answer. In the process we will see how merging fits in with the other belief change operators, and how in its case the choice perspective is a particularly apt lens. This perspective, according to which merging can be seen as a collective decision making procedure, invites the question of what tools there are to ensure that the merging process is fair. The standard set of principles used to characterize merging already include certain fairness guarantees, but they do not exhaust all the properties we would like to see instantiated. Thus, in Chapter 5 we put forward a series of novel properties, all of which address, in some way, the notion of fairness and enrich the merging landscape. Most of these properties come from the social choice literature, where they have been used to understand voting procedures [Zwicker, 2016, Baumeister and Rothe, 2016]: here is an example of social choice coming to the aid of belief change, by suggesting a series of dimensions along which to judge merging operators.

Finally, let us revisit the example we started with, of a doctor trying to decide on a course of treatment for a new disease.

Example 1.6: Treating a novel disease with novel hunches of what works

Doctor 1 from Example 1.1 is faced with the same dilemma of choosing what combination of drugs a and b to administer to sick patients. Initially, the doctor is

inclined to think that the best course of action is to use a and b together, followed by a alone, followed by b alone. Doing nothing seems like the worst thing to do, and comes last in the doctor's list of actions to take.

But after a couple of weeks of administering the cocktail of a and b drugs, the doctor realizes that they are not effective at all, and might even be harming patients. Thus, the doctor becomes convinced that administering nothing is better than administering a and b together, and revises their policy accordingly. However, this leaves a gap that puzzles the doctor: where does this leave the treatments that consist of just a and b? Does swapping the a and b pair with the option of administering nothing downgrade these treatments as well? Or do they still stand as better options than doing nothing?

We want to construe the doctor's epistemic state, in this case, as a preference over the possible treatment options. Example 1.6, then, describes a scenario when the preference order itself undergoes revision. That it makes eminent sense to revise a preference order jumps out if we see preferences as expressions not of taste, or whim, but as the result of a deliberative process: the doctor in Example 1.6 arrives at their ranking of the treatment options after weighing various factors, such as the known properties of the drugs and their own experience. As such, the preference ranking reflects the doctor's judgment of how effective the treatment options are relative to each other, and is subject to examination and revision in the same way that a judgment is. This view, according to which preferences function as *comparative evaluations* of alternatives [Hausman, 2011], makes it possible for an agent to change its preferences as it gathers more information or feedback from the external world. Example 1.6 showcases a scenario in which this is precisely what happens: if we see a preference order as made up of individual comparisons, e.g., the comparison between administering drugs a and b versus doing nothing, then change can be triggered by finding out that some of these comparisons are wrong. The agent then has to adjust its preference around this new information, keeping some of the old comparisons and potentially discarding others. What is kept and what is discarded is best construed, of course, as a matter of choice; and where there is choice there are preferences.

We will call this process *preference revision*, and in Chapter 7 we will present a set of principles and results that characterize this operation in terms of preferences over the basic comparisons that go into making a preference order. To put it more succinctly, revising preferences amounts to having preferences over preferences. We will see that the basic apparatus of belief revision lends itself to modeling preference change, but not in a straightforward manner: due to the nature of preference orders, concrete operators end up looking more like enforcement operators as described in Section 3.3 and anticipated by Example 1.4; and, as for revision of Horn formulas, extra care needs to be taken such that changes brought to the input do not alter the basic format of the epistemic state. This is, then, another example where the principles of belief change get to be applied outside their comfort zone, with the help of tools from choice theory.

What is to come

In Chapter 2 we will introduce the main background on propositional logic and rational choice that we will need for the remaining part of this work. This material is largely expository, and the results presented in it are based on standard references.

In Chapter 3 we will look at the basic models of the belief change operations that interest us: revision, update, enforcement and merging. Each operator is analyzed along a number of typical dimensions: a set of characteristic postulates; a family of characteristic preference relations on outcomes; a specific choice function that connects the two, via a representation result; and a few concrete, usually distance-based operators that fit into the outlines drawn by the postulates. The material on revision, update and merging is based on known results, and it is presented in a way that emphasizes the role of choices and preferences. The material on enforcement is part of our contribution to this thesis, and is based on work published at IJCAI 2018 [Haret et al., 2018c].

In Chapter 4 we study variants of revision that swap one of the standard postulates with alternative versions, taken to encode different biases an agent can have towards the prior information. We present postulates, preferences on outcomes that track these postulates, and show how to construct distance-based operators based on intuitive choice functions that exhibit these biases. The method we put forward manages to both capture known revision operators, and to introduce some new ones. Chapter 4 is based on work published at NMR 2018 [Haret and Woltran, 2018] and IJCAI 2019 [Haret and Woltran, 2019].

Chapter 5 looks at merging as a multi-agent social procedure, and puts forward a number of properties aimed at capturing different aspects of fairness: insensitivity to syntax, a consideration of broad lines of agreement in the profile that we call here *collective efficiency*, a sensitivity to changes in the profile that we call here *responsiveness*, strategyproofness and proportionality. These properties come in the form of postulates, and we follow common practice in mapping the postulates onto preferences on outcomes and checking existing merging operators against these postulates. Sections 5.1, 5.2 and 5.3 are based on work published at ECAI 2016 [Haret et al., 2016b], Section 5.4 is based on work published at JELIA 2019 [Haret and Wallner, 2019], and Section 5.5 is based on work published at AAAI 2020 [Haret et al., 2020].

Chapter 6 looks at revision and update for Horn formulas, for which the challenge is different than in the other chapters: the issue now is not of how to fit classical operators into new intuitions, but to find new ways of enforcing classical intuitions. This involves coming up with postulates tailored specifically for the Horn fragment, which manage to prop up conclusions that follow naturally in standard propositional logic, but fail in the Horn fragment. Since existing operators turn out to be a poor fit for the Horn fragment, new ones have to be designed. Sections 6.3 and 6.4 are based on work presented at IJCAI 2018 [Creignou et al., 2018a], while Section 6.2 is partly based on work presented at NMR 2018 [Haret and Woltran, 2018], though the representation theorem is new.

Chapter 7 puts forward a model of preference revision that follows the lines of standard

belief change operators, with postulates, preferences and choice functions that link the two. The input to a preference revision operator is a preference order, conceived of as a set of comparisons between items; what a preference revision operator does, then, is to choose which of these comparisons it will give up, if a contradiction occurs. In its details, this models is mostly similar to the model for enforcement. Chapter 7 has not been published anywhere, and can be considered new.

Conclusions and discussions of related work are offered at the end of every chapter, but Chapter 8 provides a more general overview of the material preceding it, together with some more general musings and thoughts on future work.

Finally, in a disclosure of what is not to come, we mention that there is also additional work that has been published in the lead up to this thesis, that has not made it into the material to follow, for lack of an obvious fit, but still relates to either belief change or social choice. The belief change strand includes work on merging of Horn formulas [Haret et al., 2015, Haret et al., 2017], a type of reverse merging operation we called *distribution* [Haret et al., 2016a], deviation of belief change operators with respect to fragments of propositional logic [Haret and Woltran, 2017], revision of argumentation frameworks [Diller et al., 2015, Diller et al., 2018] and merging of argumentation frameworks [Delobelle et al., 2016]. The social choice strand includes work on aggregation of incomplete CP-nets [Haret et al., 2018b] and solutions for constructing a ranking on items based on a ranking of sets of items [Haret et al., 2018a].

CHAPTER 2

Preliminaries

We generally aim to introduce notation on a need-to-know basis. However, there are some notions that permeate the entirety of this work and thus can suffer no delay.

2.1 Propositional logic

We will assume a finite set A of *propositional atoms*, intended to represent issues that can be the subject of reasoning, thought or deliberation. A *literal* l is either an atom p, in which case l is a *positive literal*, or its negation $\neg p$, in which case l is a *negative literal*. If l is a positive literal p or a negative literal $\neg p$, the *dual* $\bar{l}$ *of* l is $\neg p$ or p, respectively.

The *set* $\mathcal{L}$ *of propositional formulas* is generated from A using the usual propositional connectives ($\wedge$, $\vee$, $\neg$, $\rightarrow$ and $\leftrightarrow$), as well as the constants $\bot$ and $\top$. Propositional formulas in $\mathcal{L}$ will be used to represent either attitudes with respect to the issues in A assumed to stem from a single agent i and typically denoted by φ, or φ_i when we need to be explicit about the agent supplying the attitude, or information with respect to these issues, assumed to stem from some authoritative source and typically denoted by μ. We will want to be flexible, to a certain degree, with respect to what sort of attitude a formula φ, or φ_i, actually represents: the nominal term will be belief, i.e., conviction about what is the case, but φ can also encode opinion about what should be the case, or what is desired to be the case; in general, we will take φ to encode some constraint on the issues in A the agent is partial to, and will be explicit when using it with a more concrete meaning.

An *interpretation* w (alternatively, a *truth-value assignment*, or *outcome*) is a function mapping every atom in A to either *true* or *false*, and typically denoted by w or v. Since an interpretation w is completely determined by the set of atoms in A it makes true, we will identify w with this set of atoms and, if there is no danger of ambiguity, display w

as a word where the letters are the atoms assigned to true. The *universe* $\mathcal{U}$ is the set of all interpretations for formulas in $\mathcal{L}$.

The models of a propositional formula φ are the interpretations that satisfy it, and we write $[\varphi]$ for the set of models of φ. If φ_1 and φ_2 are propositional formulas, we say that φ_1 *entails* φ_2, written $\varphi_1 \models \varphi_2$, if $[\varphi_1] \subseteq [\varphi_2]$, and that they are *equivalent*, written $\varphi_1 \equiv \varphi_2$, if $[\varphi_1] = [\varphi_2]$. A propositional formula φ is *consistent* if $[\varphi] \neq \emptyset$ and *refutable* if $[\varphi] \neq \mathcal{U}$. The models of $\bot$ and $\top$ are $[\bot] = \emptyset$ and $[\top] = \mathcal{U}$, i.e., $\bot$ has no model and $\top$ is satisfied by every interpretation in the universe. A propositional formula is *complete* if it has exactly one model. Complete formulas are sometimes denoted by $\dot{\varphi}$, with the dot suggesting that $\dot{\varphi}$ has one model. The sets of consistent, refutable and complete formulas are $\mathcal{L}_{\text{con}}$, $\mathcal{L}_{\text{ref}}$ and $\mathcal{L}_{\text{comp}}$, respectively.

Example 2.1: Propositional formulas and their semantics

The scenario in Example 1.2 can be modeled by using propositional variables to represent the issues under contention: that humans use tools (a), that chimpanzees are part of a different species from humans (b) and that chimpanzees use tools (c). Thus, the set of atoms is $A = \{a, b, c\}$, with the universe, i.e., the set of all possible interpretations, being $\mathcal{U} = \{\emptyset, a, b, c, ab, ac, bc, abc\}$.

One of the prevailing beliefs of the primatology community, circa 1960, can be taken to be that humans were the only species capable of using tools, which we can approximate here with the implication that if humans use tools and chimpanzees are different from humans, then there is no way that chimpanzees can use tools, represented by the propositional formula $\varphi_1 = (a \wedge b) \rightarrow \neg c$ whose set of models is $[\varphi_1] = \{\emptyset, a, b, c, ab, ac, bc\}$. Note that the interpretations where a and c are true and b is false, i.e., the outcome according to which humans and chimpanzees use tools and chimpanzees are the same species as humans, is written as ac. It is straightforward to see that φ_1 is both consistent and refutable, but not complete.

To φ_1 we can add the common-sense beliefs that humans use tools and that they are truly different from humans, i.e., the formula $\varphi_2 = a \wedge b$, with $[\varphi_2] = \{ab, abc\}$, to get a snapshot of the ensemble of ideas that Jane Goodall eventually proved wrong. This ensemble can be represented as the conjunction of the formulas φ_1 and φ_2, i.e., the propositional formula $\varphi = \varphi_1 \wedge \varphi_2 = (a \wedge b) \wedge ((a \wedge b) \rightarrow \neg c)$. It holds that $[\varphi] = \{ab\}$, and hence φ is complete.

There are certain transformations we will want to subject propositional formulas to in order to set limits for belief change operators: they include replacing literals with their duals, changing the order of elements in a tuple and renaming atoms. These transformations are usually done to the syntax of the formulas, but they also have effects on the semantic level, and we must dedicate a few paragraphs to charting out these effects. Thus, if φ is a propositional formula, the *dual* $\overline{\varphi}$ *of* φ is a formula obtained by replacing every literal l in φ with its dual $\bar{l}$. A similar operation can be defined on the

semantic side: if w is an interpretation, the *dual* $\overline{w}$ *of* w is the complement of w, i.e, $\overline{w} = A \setminus w$. If $\mathcal{W}$ is a set of interpretations, the *dual* $\overline{\mathcal{W}}$ *of* $\mathcal{W}$ is the set of interpretations containing the duals of the interpretations in $\mathcal{W}$, i.e., $\overline{\mathcal{W}} = \{\overline{w} \mid w \in \mathcal{W}\}$.

Example 2.2: Duals of formulas and their models

For the set of atoms $A = \{a, b, c\}$ and the propositional formula $\varphi_1 = (a \wedge b) \rightarrow \neg c$, then the dual of φ_1 is the formula $\overline{\varphi_1} = (\overline{a} \wedge \overline{b}) \rightarrow \overline{\neg c} = (\neg a \wedge \neg b) \rightarrow c$. The set of models of φ_1 is $[\varphi_1] = \{\emptyset, a, b, c, ab, ac, bc\}$, with the duals of interpretation $\emptyset$ and a being $\overline{\emptyset} = abc$ and $\overline{a} = bc$. Note that $\overline{[\varphi_1]} = \{abc, bc, ac, ab, c, b, a\}$.

In Example 2.2 it is the case that models of the dual $\overline{\varphi}$ of φ are exactly the duals of the models of φ. Though we do not provide a formal proof, we mention here that this holds more generally.

Proposition 2.1

If φ is a propositional formula, then $[\overline{\varphi}] = \overline{[\varphi]}$.

We will use the notion of a dual of a formula φ in Chapters 4 and 5, to encode something like the point of view diametrically opposed to φ. Throughout it all, Proposition 2.1 will repeatedly come in handy.

If $N = \{1, \ldots, n\}$ is a set of integers, a *permutation* σ *of* N is the familiar notion of a bijective function $\sigma\colon N \rightarrow N$. If σ is a permutation of $N = \{1, \ldots, n\}$, then *the inverse* σ^{-1} *of* σ is the bijection $\sigma^{-1}\colon N \rightarrow N$ such that $\sigma^{-1}(\sigma(i)) = i$, for any $i \in N$, i.e., the permutation that reverses σ. We will typically use permutations in a context in which $N = \{1, \ldots, n\}$ is a set of agents, each with their own opinion φ_i, and $(\varphi_i)_{1 \leq i \leq n}$ is the n-tuple that contains their opinions. In this context, $(\varphi_{\sigma(i)})_{1 \leq i \leq n}$ is the tuple that swaps the order of the agents around.

On occasion we will also want to swap atoms in A around: technically, if atoms in A are indexed by an integer, e.g., a_1, a_2, ..., this can be achieved simply by a permutation of the indices. But since, for presentation purposes, we usually denote atoms with distinct letters, we introduce a special notion, called a renaming of the atoms in A. Formally, a *renaming* ρ *of* A is exactly what we expect it to be, i.e., a bijective function $\rho\colon A \rightarrow A$. The *inverse* ρ^{-1} *of* ρ is a permutation such that $\rho^{-1}(\rho(p)) = p$, for any atom $p \in A$. If φ is a propositional formula, the *renaming* $\rho(\varphi)$ *of* φ is a formula $\rho(\varphi)$ whose atoms are replaced according to ρ. On the semantic side, if w is an interpretation and ρ is a renaming of A, the *renaming* $\rho(w)$ *of* w is an interpretation obtained by replacing every atom p in w with $\rho(p)$. If $\mathcal{W}$ is a set of interpretations, the *renaming* $\rho(\mathcal{W})$ *of* $\mathcal{W}$ is defined as $\rho(\mathcal{W}) = \{\rho(w) \mid w \in \mathcal{W}\}$, i.e., the set of interpretations whose elements are the renamed interpretations in $\mathcal{W}$.

Example 2.3: Permutations and renamings

For the set $N = \{1, 2, 3, 4\}$ of Academy members in Example 1.5, consider the permutation σ according to which $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 4$ and $\sigma(4) = 1$. If $(\varphi_1, \varphi_2, \varphi_3, \varphi_4)$ is a tuple consisting of their opinions, then applying the permutation σ to N results in $(\varphi_2, \varphi_3, \varphi_4, \varphi_1)$

If the set of atoms is $A = \{a, b, c\}$, consider a renaming ρ such that $\rho(a) = b$, $\rho(b) = c$ and $\rho(c) = a$. If a, b and c stand for the directors Alma Har'el, Bong Joon Ho and Céline Sciamma, respectively, from in Example 1.5, then the first Academy member's opinion can be represented by the propositional formula $\varphi_1 = a \wedge b$. Applying the renaming ρ gives us that $\rho(\varphi_1) = (\rho(a) \wedge \rho(b)) = (b \wedge c)$. On the semantic side, we have that $[\varphi_1] = \{ab, abc\}$ and $\rho([\varphi_1]) = \{\rho(ab), \rho(abc)\} = \{bc, abc\}$.

In Example 2.3 it holds that the set of models of a renamed formula φ is the same as the set of renamed models of φ. This, also, holds more generally.

Proposition 2.2

If ρ is a renaming of the set A of atoms and φ is a propositional formula, then $[\rho(\varphi)] = \rho([\varphi])$.

Another thing we will be eminently interested in is the relationship between a propositional formula and its semantics, consisting of sets of interpretations: primarily, the assurance that we can move freely between the two. This is done through the notion of proxy formulas. Thus, if $\mathcal{W} = \{w_1, \ldots, w_k\}$ is a set of interpretations, an *$\mathcal{L}$-proxy $\varepsilon_{\mathcal{W}}$ of $\mathcal{W}$* is a propositional formula such that $[\varepsilon_{\mathcal{W}}] = \{w_1, \ldots, w_k\}$. At the same time, an *$\mathcal{L}$-antiproxy $\varepsilon_{-\mathcal{W}}$ of $\mathcal{W}$* is a propositional formula $\varepsilon_{-\mathcal{W}}$ such that $[\varepsilon_{-\mathcal{W}}] = \mathcal{U} \setminus \{w_1, \ldots, w_k\}$. We will want to refer to proxies and antiproxies through various shorthands. Thus, if there is no danger of ambiguity, we write $\varepsilon_{w_1,\ldots,w_k}$, or even more simply, $\varepsilon_{1,\ldots,k}$, and $\varepsilon_{-w_1,\ldots,-w_k}$, or $\varepsilon_{-1,\ldots,-k}$, instead of $\varepsilon_{\mathcal{W}}$ and $\varepsilon_{-\mathcal{W}}$, respectively. Intuitively, an $\mathcal{L}$-proxy of a set $\mathcal{W}$ of interpretations is a propositional formula that encodes, possibly in a compact way, all the outcomes in $\mathcal{W}$, while an $\mathcal{L}$-antiproxy is a propositional formula that encodes the complement of $\mathcal{W}$.

Example 2.4: Proxies and antiproxies

If $A = \{a, b\}$ is the set of atoms, an $\mathcal{L}$-proxy of the set of interpretations $\mathcal{W} = \{\emptyset, ab\}$ is the disjunctive normal form (DNF) formula $\varphi_1 = (\neg a \wedge \neg b) \vee (a \wedge b)$. Note, φ_1 is not the only $\mathcal{L}$-proxy of a and b; $\varphi_2 = a \leftrightarrow b$ works just as well. An $\mathcal{L}$-antiproxy of $\mathcal{W}$ is a formula whose set of models is $\mathcal{U} \setminus \mathcal{W} = \{a, b\}$, examples of which are $\varphi_3 = (a \wedge \neg b) \vee (\neg a \wedge b)$ or $\varphi_4 = (a \leftrightarrow \neg b)$.

As Example 2.4 makes clear, $\mathcal{L}$-proxies (and $\mathcal{L}$-antiproxies) of sets of interpretations

always exist (e.g., as DNF formulas), but are not necessarily unique. For our purposes, existence is much more important than uniqueness. Since we will typically try to abstract away as much as possible from the syntax of formulas, non-uniqueness of a proxy formula will usually not be a factor in the results to follow.

In a multi-agent scenario we assume a set $N = \{1, \ldots, n\}$ of n agents. An *$\mathcal{L}$-profile* $\vec{\varphi}$ (alternatively, a *propositional profile* $\vec{\varphi}$) is an n-tuple $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$ of propositional formulas, also written as $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$, where each formula φ_i is assumed to correspond to an agent i. The set of all propositional profiles is $\mathcal{L}^n$. As for the single-agent case, we want to be liberal with respect to the meaning assigned to φ_i: it can represent agent i's belief, preference, judgments, goals or knowledge. The *models* $[\vec{\varphi}]$ *of a propositional profile* $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$ are the interpretations satisfying all formulas in $\vec{\varphi}$, i.e., $[\vec{\varphi}] = \bigcap_{i=1}^{n} [\varphi_i]$. As for a propositional formula, a propositional profile $\vec{\varphi}$ is *consistent* if $[\vec{\varphi}] \neq \emptyset$. A propositional profile is *complete* if every formula in it is complete. The sets of consistent and complete profiles are $\mathcal{L}^n_{\text{con}}$ and $\mathcal{L}^n_{\text{comp}}$, respectively. If $\vec{\varphi}_1 = (\varphi_i)_{1 \leq i \leq n}$ and $\vec{\varphi}_2 = (\varphi_i)_{n+1 \leq i \leq p}$ are profiles, $\vec{\varphi}_1 + \vec{\varphi}_2$ is the profile $\vec{\varphi}_1 + \vec{\varphi}_2 = (\varphi_i)_{1 \leq i \leq p}$ obtained by appending $\vec{\varphi}_2$ to $\vec{\varphi}_1$. If φ is a formula and there is no danger of ambiguity, we write $\vec{\varphi} + \varphi$ instead of $\vec{\varphi} + (\varphi)$. Two profiles $\vec{\varphi}_1$ and $\vec{\varphi}_2$ are *equivalent* if there exists a bijection $f : \vec{\varphi}_1 \to \vec{\varphi}_2$ such that, for any $\varphi_i \in \vec{\varphi}_1$, it holds that $\varphi_i \equiv f(\varphi_i)$.

Example 2.5: Propositional profiles

The scenario in Example 1.5 can be modeled by using propositional variables to stand for the best director candidates: Alma Har'el (a), Bong Joon Ho (b) and Céline Sciamma (c). The opinions of the four Academy members can be represented by four propositional formulas φ_1, φ_2, φ_3 and φ_4, with $\varphi_1 = a \wedge b$, $\varphi_2 = a \wedge (b \vee c)$, $\varphi_3 = \neg a \wedge b \wedge \neg c$, $\varphi_4 = \neg a \wedge \neg b \wedge c$. The profile consisting of the (opinions of) the first three Academy members is $\vec{\varphi} = (\varphi_1, \varphi_2, \varphi_3)$. If we append the fourth member, the profile is $\vec{\varphi} + \varphi_4 = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$.

2.2 Preferences: preorders, partial and total

We will typically use binary relations $\leq$ on a set X of alternatives to encode some kind of preference (or priority, or plausibility) relation over the elements of X, with $\leq$ being typically referred to as a *preference relation on* X and $x \leq x'$ to be read as saying that x *is at least as good (or important, or plausible) as* x' *with respect to* $\leq$. If $\leq$ is a preference relation on X and x and x' are two alternatives in X, then x *is strictly better than* x' *with respect to* $\leq$, written $x < x'$, if $x \leq x'$ and $x' \not\leq x$; x *and* x' *are indifferent with respect to* $\leq$, written $x \approx x'$, if $x \approx x'$ and $x' \approx x$; finally, x *and* x' *are incomparable with respect to* $\leq$ if $x \not\leq x'$ and $x' \not\leq x$. Intuitively, if two alternatives x and x' are indifferent with respect to a preference order $\leq$, this is like saying that there is nothing to set them apart, and they are equally good. If x and x' are incomparable with respect to $\leq$, this may be because it is not known how x and x' fare with respect to each other, or because,

for whatever reason, there is simply no fact of the matter either way. If $\leq$ is a preference order on X, then the *transitive closure* $\leq^+$ *of* $\leq$ is defined, for any alternatives x and x' in X, as:

$$x \leq^+ x' \text{ if there exist } x_1, \ldots, x_k \text{ in } X \text{ such that } x \leq x_1 \leq \cdots \leq x_k \leq x'.$$

Clearly, if $x \leq x'$, then it also holds that $x \leq^+ x'$, since we can just take $x_1 = x$ and $x_k = x'$. Thus, $\leq^+$ contains all the comparisons in $\leq$, together with the comparisons inferred using an intermediary chain of comparisons.

If $\leq$ is a preorder on X, the $\leq$-*minimal elements* $\min_{\leq} X$ and $\leq$-*maximal elements* $\max_{\leq} X$ of X are defined, respectively, as:

$$\min_{\leq} X \stackrel{\text{def}}{=} \{x \in X \mid \text{there is no } x' \in X \text{ such that } x' < x\},$$
$$\max_{\leq} X \stackrel{\text{def}}{=} \{x \in X \mid \text{there is no } x' \in X \text{ such that } x' > x\}.$$

As per our convention regarding the meaning of $\leq$, x being $\leq$-minimal in X means that there is no other element in X strictly better than x and, similarly, x being $\leq$-maximal in X means that there is no other element in X strictly worse than x.

In order to function as a preference over the elements of X, a binary relation $\leq$ is expected to satisfy, for any integer n and alternatives x, x_1, …, x_n in X, some selection of the following properties:

(Pr_1) $x \leq x$. (reflexivity)

(Pr_2) If $x_1 \leq x_2$ and $x_2 \leq x_3$, then $x_1 \leq x_3$. (transitivity)

(Pr_3) If $x_1 \neq x_2$, then $x_1 \leq x_2$ or $x_2 \leq x_1$. (totality)

(Pr_{SC}) If $x_1 \leq \cdots \leq x_n$, then $x_n \not\leq x_1$ (Suzumura consistency)

If $\leq$ is a binary relation on a set X of alternatives, then $\leq$ is a *preorder on* X if $\leq$ satisfies properties Pr_{1-2}, i.e., if $\leq$ is reflexive and transitive. We write $\mathcal{P}_X$ for the set of preorders on X. If $\leq$ is a preorder on X, then $\leq$ is *total* if it also satisfies property Pr_3, i.e., if any two distinct alternatives in X are the subject of some comparison in X, and we write $\mathcal{T}_X$ for the set of total preorders on X. Note that the transitive $\leq^+$ closure of any preference order $\leq$ is, by definition, transitive.

Example 2.6: Preorders

Consider a set of alternatives $X = \{x_1, x_2, x_3, x_4, x_5\}$ and twe preorders, $\leq_1$ and $\leq_2$, on X, depicted in Figure 2.1. Of these preorders, $\leq_1$ is partial and $\leq_2$ is total. We have that $\min_{\leq_1} X = \min_{\leq_2} X = \{x_1\}$. However, if we consider the restriction of $\leq_1$ and $\leq_2$ to the set $X' = \{x_3, x_4, x_5\}$, then $\min_{\leq_1} X' = \{x_3, x_4\}$ and $\min_{\leq_2} X' = \{x_4\}$.

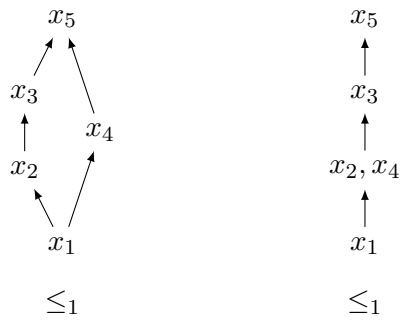


Figure 2.1: A partial preorder $\leq_1$ and a total preorder $\leq_2$ on the set $X = \{x_1, x_2, x_3, x_4, x_5\}$ of alternatives. An arrow from x_i to x_j means that x_i is strictly better than x_j, such that better alternatives are depicted lower; if x_i and x_j are separated by a comma, that means they are indifferent; and if x_i and x_j are drawn apart, then they are incomparable. In the interest of readability, arrows inferred by transitivity are ommitted: nonetheless, since both $\leq_1$ and $\leq_2$ are assumed to be preorders, then they must be understood to be transitive, even when the corresponding arrows are absent.

Note that if $\leq$ is a total preorder on X, then the $\leq$-minimal elements in X end up being the overall best elements in X, i.e., if $x \in \min_{\leq} X$, then $x \leq x'$, for any $x' \in X$. As Example 2.6 illustrates, this does not hold if $\leq$ is a partial preorder.

Preorders are a key ingredient in traditional belief change, and we will make use of both partial and total preorders on interpretations (i.e., in which the set of alternatives is the universe $\mathcal{U}$) in Chapters 3, 4, 5 and 6. A different type of preference order will be used in Chapter 7, but we defer definitions for those until we need them.

Property Pr_{SC}, where 'SC' stands for *Suzumura consistency* [Suzumura, 1976, Suzumura, 1983, Bossert and Suzumura, 2010], says that it is not possible to form a chain of comparisons that starts with x_1 and ends with x_n, in which every alternative is at least as good as the next one, but the last one, x_n ends up being strictly better than the first one, x_1. A preference order $\leq$ on a set of alternatives X is *Suzumura consistent* if it satisfies property Pr_{SC}. Suzumura consistency is a weakening of the transitivity property Pr_2: clearly, any relation $\leq$ that is transitive is also Suzumura consistent, i.e., property Pr_2 implies property Pr_{SC}; the converse, however, does not hold.

Example 2.7: Suzumura consistency does not imply transitivity

For the set of alternatives $X = \{x_1, x_2, x_3\}$, take a preference order $\leq$ such that $x_1 \leq x_2$ and $x_2 \leq x_3$, but x_1 and x_3 are incomparable. Clearly, $\leq$ is Suzumura consistent but not transitive.

Suzumura consistency is of interest to us because the rational choice literature has identified it as one of the safest fallback options when some preference information is available that can be pieced together into a preference order $\leq$, but, for whatever reason,

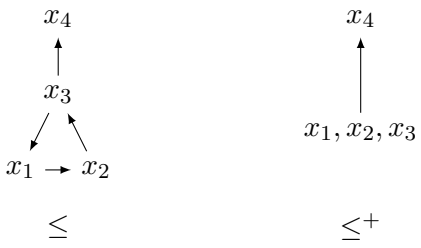


Figure 2.2: Preference relation $\leq$ that does not admit of an ordering extension: the transitive closure $\leq^+$ of $\leq$ does not preserve the strict comparisons of $\leq$.

$\leq$ is not transitive. Since transitivity is, in general, a desirable property of any preference order, it would be good if a transitive order could be constructed on the back of $\leq$, and the obvious suggestion is to replace $\leq$ with its transitive clousre $\leq^+$. However, there are cases in which the transitive closure would flatten much preference information that we would like to see preserved, and thus lead to an undesirable result. To make this more precise, we can bring in the notion of an ordering extension. If X is a set of alternatives and $\leq$ and $\leq'$ are binary relations on X, then $\leq'$ *extends* $\leq$ *on* X if, for any alternatives x_1 and x_2 in X, the following properties hold:

(i) if $x_1 \leq x_2$, then $x_1 \leq' x_2$;

(ii) if $x_1 < x_2$, then $x_1 <' x_2$.

Intuitively, $\leq'$ extends $\leq$ on X if $\leq$ contains all the comparisons in $\leq$ and, additionally, $\leq'$ contains all the strict comparisons of $\leq$: presumably, the strict comparisons are hard won pieces of information that we would not want to lose. Furtermore, $\leq'$ is an *ordering extension of* $\leq$ *on* X if $\leq'$ extends $\leq$ and $\leq'$ satisfies properties Pr_{1-3}, i.e., if $\leq'$ is a total preorder on X that extends $\leq$.

Example 2.8: Ordinal extensions

Note, first, that $\leq_2$ in Example 2.6 extends $\leq_1$ and, what is more, $\leq_2$ is an ordering extension of $\leq_1$, since it preserves all the strict comparisons of $\leq_1$.

Consider, now, a set of alternatives $X = \{x_1, x_2, x_3, x_4\}$ and the preference relation $\leq$ on X depicted on the left in Figure 2.2, where $x_1 < x_2 < x_3 < x_1$, $x_2 \not< x_1$, $x_3 \not< x_2$, $x_1 \not< x_3$, and $x_i < x_4$, for $i \in \{1, 2, 3\}$. Note that $\leq$ is not transitive, and is thus problematic. However, the transitive closure $\leq^+$ of $\leq$, depicted on the right in Figure 2.2, is arguably also problematic, as it flattens the cycle between x_1, x_2 and x_3 and does not preserve the strict comparisons in $\leq$. The transitive closure $\leq^+$ of $\leq$, then, is not an ordering extension of $\leq$. In fact, it is easy to see that $\leq$ does not admit an ordering extension. Incidentally, $\leq$ is not Suzumura consistent either.

The primary question about ordering extensions concerns their existence: ideally, we would like the preference relation $\leq$ we are working with to be reflexive and transitive, i.e., a preorder, at the least. But there are situations, and we will encounter them in Chapters 6, where these properties cannot be guaranteed, and a much weaker relation has to be contended with. The only hope, in this situation, is to massage $\leq$ into a more manageable format: if $\leq$ can be extended, in a meaningful way, to a total preorder, then it can still fulfil its assigned role as the basis for a decision procedure. We would like to know, then, exactly how weak $\leq$ can be such that it can still guide a rational decision maker in its choices. The answer turns out to hinge on Suzumura consistency. Indeed, ensuring that a preference relation $\leq$ is Suzumura consistent turns out to both a sufficient and a necessary condition for the possibility of extending the relation to a total preorder.

Theorem 2.1 ([Suzumura, 1976])

If X is a (potentially infinite) set of alternatives and $\leq$ is a binary relation on X, then there exists an ordering extension $\leq'$ of $\leq$ on X if and only if $\leq$ is Suzumura consistent.

Theorem 2.1 extends an earlier result that provided only a sufficient condition for the existence of an ordering extension [Szpilrajn, 1930], and applies to both finite and infinite sets of alternatives, though in this work we will mainly be concerned with finite sets of alternatives.

2.3 Distances and aggregation functions

The primary devices for generating preorders (either total or partial) on interpretations we make recourse to in this work are a *dissimilarity function d between interpretations*, used to quantify the disagreement between two outcomes, and an *aggregation function* $\oplus$, used to boil down vectors $(x_i)_{1\leq i\leq n}$ of dissimilarity measures to forms that can be meaningfully compared to each other.

Distances

Formally, a dissimilarity function d between interpretations is a function $d\colon \mathcal{U}\times\mathcal{U}\to\mathbb{R}_{\geq 0}$ expected to satisfy, for any interpretations w, w_1, w_2 and w_3, some subset of the following properties:

(D$_1$) $d(w,w)=0$. (identity of indiscernibles)

(D$_2$) If $w_1\neq w_2$, then $d(w_1,w_2)>0$. (non-negativity)

(D$_3$) $d(w_1,w_2)=d(w_2,w_1)$. (symmetry)

(D$_4$) $d(w_1,w_3)\leq d(w_1,w_2)+d(w_2,w_3)$. (triangle inequality)

There is some variation within the literature with respect to what is called how, but in the spirit of standard references [Deza and Deza, 2016] we will say that a dissimilarity function d is a *quasi-distance* if it satisfies properties D_{1-2}, a *distance* if it satisfies properties D_{1-3} and a *metric* if it satisfies properties D_{1-4}. In general, we will require a dissimilarity measure d to satisfy at least properties D_{1-2}: for revision, update and enforcement a quasi-distance will be enough, whereas for merging we will require d to be a distance. Intuitively, $d(w_1, w_2)$ is supposed to measure how different w_2 is from w_1, which, in turn, becomes a proxy for how likely (or plausible, or desirable) w_2 is from the point of view of w_1. Consequently, smaller dissimilarity translates, straightforwardly enough, into a higher degree of similarity and, implicitly, into a higher degree of likelihood (or plausibility, or desirability) of w_2 relative to w_1.

Popular choices of distances between interpretations are the *Hamming distance* d_{H} and the *drastic distance* d_{D}, defined, for any interpretations w_1 and w_2, as:

$$d_{\mathrm{H}}(w_1, w_2) = |w_1 \setminus w_2| \cup |w_2 \setminus w_1|, \qquad d_{\mathrm{D}}(w_1, w_2) = \begin{cases} 0, & \text{if } w_1 = w_2, \\ 1, & \text{otherwise.} \end{cases}$$

The Hamming distance d_{H} counts the number of atoms that w_1 and w_2 differ on, while the drastic distance is much coarser, keeping track only of whether w_1 and w_2 are different or not. Both the Hamming and the drastic distances satisfy all properties D_{1-4}, so technically they are metrics, though for our purposes we will rarely make use of all these properties.

Example 2.9: Hamming and drastic distances

For the intepretations ab and ac, we have that $d_{\mathrm{H}}(ab, ac) = 2$ and $d_{\mathrm{D}}(ab, ac) = 1$.

Aggregation functions

We will need a way of talking about distances not just between individual interpretations, but also between formulas, or sets of interpretations, and interpretations, and this will involve the use of an *aggregation function* $\oplus$, which for our purposes is a function $\oplus\colon \mathbb{R}^n \to \mathbb{R}^m$ that, for integers m and n, maps n-tuples $\vec{x} = (x_i)_{1 \leq i \leq n}$ of real numbers to m-tuples $(x'_i)_{1 \leq i \leq m}$ of real numbers. This definition is a bit of an overkill, as we will only make use of the cases when $m = 1$ and $m = n$. In the case when $m = 1$, the aggregated value $\oplus\vec{x}$ is a vector containing only one value, in which case we write simply x instead of (x).

What this general definition allows us to do is to use one single method for comparing the various types of aggregated values we make use of. This method relies on the *the lexicographic order* $\leq_{\mathrm{lex}}$ *on* $\mathbb{R}^n$, defined for any integer n and n-tuples $\vec{x} = (x_i)_{1 \leq i \leq n}$ and

$\vec{y} = (y_i)_{1 \leq i \leq n}$ in $\mathbb{R}^n$, as follows:

$$\begin{aligned} \vec{x} \leq_{\text{lex}} \vec{y} \text{ if } & x_1 \leq y_1, \text{ or} \\ & \text{if } x_1 = y_1 \text{ and } x_2 \leq y_2, \text{ or} \\ & \dots \\ & \text{if } x_1 = y_1, \dots, x_{n-1} = y_{n-1} \text{ and } x_n \leq y_n. \end{aligned}$$

Note that when $n = 1$, i.e., the aggregated values of $\oplus\vec{x}$ and $\oplus\vec{y}$ are $\oplus\vec{x} = x_1$ and $\oplus\vec{y} = y_1$, for some real numbers x_1 and y_1, then comparing $\oplus\vec{x}$ and $\oplus\vec{y}$ according to $\leq_{\text{lex}}$ reduces to comparing x_1 and y_1, i.e., $\oplus\vec{x} \leq_{\text{lex}} \oplus\vec{y}$ if $x_1 \leq y_1$. In this case, we simply write that $\oplus\vec{x} \leq \oplus\vec{y}$ instead of $\oplus\vec{x} \leq_{\text{lex}} \oplus\vec{y}$.

The concrete aggregation functions of immediate interest are the min, max, leximax, leximin and sum aggregation functions, defined, for any n-tuple $(x_i)_{1 \leq i \leq n}$ and permutation σ of $\{1, \dots, n\}$, as follows:

$$\begin{aligned} \min(x_i)_{1 \leq i \leq n} &= x_i, \text{ where } x_i \leq x_j, \text{ for any } x_j \text{ in } \vec{x}, \\ \max(x_i)_{1 \leq i \leq n} &= x_i, \text{ where } x_i \geq x_j, \text{ for any } x_j \text{ in } \vec{x}, \\ \text{leximax}(x_i)_{1 \leq i \leq n} &= (x_{\sigma(i)})_{1 \leq i \leq n}, \text{ where } x_{\sigma(1)} \geq \dots \geq x_{\sigma(n)}, \\ \text{leximin}(x_i)_{1 \leq i \leq n} &= (x_{\sigma(i)})_{1 \leq i \leq n}, \text{ where } x_{\sigma(1)} \leq \dots \leq x_{\sigma(n)}, \\ \text{sum}(x_i)_{1 \leq i \leq n} &= x_1 + \dots + x_n. \end{aligned}$$

Intuitively, the min, max, leximax, leximin and sum aggregation functions return, respectively, the minimal value in $(x_i)_{1 \leq i \leq n}$, the maximal value in $(x_i)_{1 \leq i \leq n}$, $(x_i)_{1 \leq i \leq n}$ ordered in descending order, $(x_i)_{1 \leq i \leq n}$ ordered in ascending order and the sum of the values in $(x_i)_{1 \leq i \leq n}$.

Example 2.10: Aggregation functions

If $\vec{x} = (1, 0, 2)$ and $\vec{y} = (1, 1, 1)$, we have that $\min(\vec{x}) = 0$, $\max(\vec{x}) = 2$, $\text{leximax}(\vec{x}) = (2, 1, 0)$, $\text{leximin}(\vec{y}) = (0, 1, 2)$ and $\text{sum}(\vec{x}) = 3$. Thus, it holds that $\min(\vec{x}) < \min(\vec{y})$, $\max(\vec{y}) < \max(\vec{y})$, $\text{leximax}(\vec{y}) <_{\text{lex}} \text{leximax}(\vec{x})$ $\text{leximin}(\vec{x}) <_{\text{lex}} \text{leximin}(\vec{y})$ and $\text{sum}(\vec{x}) = \text{sum}(\vec{y})$.

The min aggregation function is a ubiquitous presence in belief change, and we will make use of it in all subsequent chapters. The other aggregation functions will show up in Chapter 4, alongside two other aggregation functions whose definition is deferred until needed. Functions leximax, leximin and sum are also used in merging, and will show up in Chapters 3 and 5. In the context of merging, it is usually useful that the aggregation functions satisfies a number of desirable properties. Before we spell them out, we mention that $\vec{0}$ is the tuple whose entries are uniformly 0, i.e., $\vec{0} = (0, \dots, 0)$. The properties we are interested in are, for any $x_1, \dots, x_n$ and x'_n in $\mathbb{R}$, as follows:

(Ag_1) $\oplus(x_1) = x_1$. (identity)

	a	ac	$d_{\mathrm{H}}^{\min}(\varphi, \bullet)$	$d_{\mathrm{H}}^{\max}(\varphi, \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximax}}(\varphi, \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximin}}(\varphi, \bullet)$	$d_{\mathrm{H}}^{\mathrm{sum}}(\varphi, \bullet)$
bc	3	2	2	3	$(3, 2)$	$(2, 3)$	5

Table 2.1: Hamming distances from each model of φ to bc, together with the aggregated distance $d_{\mathrm{H}}^{\oplus}(\varphi, bc)$, for the aggregation functions introduced so far.

(Ag_2) $\oplus(x_1, \ldots, x_m) = \vec{0}$ if and only if $x_i = 0$, for all $1 \leq i \leq m$. (minimality)

(Ag_3) If $x_i \leq x_i'$, then $\oplus(x_1, \ldots, x_i, \ldots, x_m) \leq_{\mathrm{lex}} \oplus(x_1, \ldots, x_i', \ldots, x_m)$. (monotonicity)

It is straightforward to see that the leximax, leximin and sum aggregation functions all satisfy properties Ag_{1-3}.

Putting dissimilarity and aggregation functions together

The main thing we want to do with dissimilarity and aggregation functions is to measure the dissimilarity between a formula φ and an interpretation w. The idea is to define this measure as the aggregate value of the dissimilarity between each model of φ and w. For the following definitions we will assume that d is a quasi-distance function, i.e., that it satisfies at least properties D_{1-2}.

If d is a quasi-distance function between interpretations, w is an interpretation, φ is a consistent propositional formula, and $\oplus$ is an aggregation function, then the $(d, \oplus)$-*induced distance* $d^{\oplus}(\varphi, w)$ *from* φ *to* w is defined as:

$$d^{\oplus}(\varphi, w) = \oplus(d(v, w))_{v \in [\varphi]}.$$

If φ is inconsistent, i.e., $[\varphi] = \emptyset$, then we establish, by convention, that $d(\varphi, w) = 0$, for any interpretation w. Intuitively, the $(d, \oplus)$-induced distance from φ to w puts a number on how close w is to φ. This number is obtained by aggregating the distances between each model of φ and w using the quasi-distance function d and the aggregation function $\oplus$. This will allow us to compare intepretations with respect to each other, relative to a formula φ, when needed.

Example 2.11: Keeping up with the humans

The scenario in Example 1.3 can be modeled using propositional variables to represent the indicators my smarthome keeps track of: whether the temperature inside the house is above 15° C (a), whether the Wi-Fi is on after 21:00 (b), and whether my friend is online after 21:00 (c). Thus, the set of atoms is $A = \{a, b, c\}$. My smarthome is set up to make sure that a is true and that b is not, which we can represent as the propositional formula $\varphi = a \wedge \neg b$, with $[\varphi] = \{a, ac\}$. Consider, now, the interpretation bc. We have that $d_{\mathrm{H}}(a, bc) = 3$, since a and bc differ with respect to three atoms, whereas $d_{\mathrm{D}}(a, bc) = 1$, since a and bc are different. The vector of Hamming distances

from φ to bc is $(d_{\mathrm{H}}(v, bc))_{v\in[\varphi]} = (d_{\mathrm{H}}(a, bc), d_{\mathrm{H}}(ac, bc)) = (3, 2)$. Thus, the distances from φ to bc, using the aggregation functions introduced so far, are as follows: $d_{\mathrm{H}}^{\min}(\varphi, bc) = 2$, $d_{\mathrm{H}}^{\max}(\varphi, bc) = 3$, $d_{\mathrm{H}}^{\mathrm{leximax}}(\varphi, bc) = (3, 2)$, $d_{\mathrm{H}}^{\mathrm{leximin}}(\varphi, bc) = (2, 3)$ and $d_{\mathrm{H}}^{\mathrm{sum}}(\varphi, bc) = 5$. The distances are also depicted in Table 2.1.

We will use $(d, \oplus)$-induced distances virtually throughout the entire thesis, whenever in need of a constructive way of ranking outcomes relative to a particular formula φ. In Section 3.4 we will go even further and define the distance between a profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$ and an interpretation w: aggregation functions will make another appearance there.

2.4 Rational choice, individual and social

Rational choice theory is a vast topic, and our purpose is not to do it full justice here, beyond mentioning the basic idea at its core. This idea is that decision-makers have coherent preferences and choose the best alternatives from the ones available. The theoretical part that is of immediate relevance to us concerns the axioms guiding rational choice functions and the way in which they tie in with preferences. We will look at the single-agent and multi-agent cases separately.

For the purposes of this section, we will assume a finite set X of alternatives. The set 2^X is the set of subsets of X, and, for an integer k, 2^X_k is the set of subsets of X of size k.

Individual choice

If X is a set of alternatives, a *choice function* c is a function $c\colon 2^X \to 2^X$ that takes as input a subset M of X, called a *choice set*, or *menu*, and returns a subset $c(M)$ of X, called the *chosen alternatives*. Intuitively, the choice function c models the behavior of an agent confronted with a range of alternatives, from which some are chosen. The agent can choose whatever it wants from the choice set, as long as it does so in a way that respects some basic standards of rationality. Traditionally, a rational choice function c is expected to satisfy, for any choice sets M, M_1 and M_2, the following properties:

(C_1) $c(M) \subseteq M$.

(C_2) If $M \neq \emptyset$, then $c(M) \neq \emptyset$.

(C_3) If $M_2 \subseteq M_1$, then $c(M_1) \cap M_2 \subseteq c(M_2)$.

(C_4) If $M_2 \subseteq M_1$ and $x_1, x_2 \in c(M_2)$, then $x_1 \in c(M_1)$ if and only if $x_2 \in c(M_1)$.

Property C_1 says that the elements chosen from a choice set M should be, as a matter of fact, elements of M, and is intended to be as uncontroversial as it sounds. Property C_2 says that if the choice set M is non-empty, i.e., there is something to choose from, then the choice $c(M)$ is non-empty, i.e., something is chosen. Property C_3, sometimes called

property α [Sen, 1969, Sen, 1970], says that if $M_2 \subseteq M_1$, then any elements chosen from M_1 that also happen to be in M_2 are also chosen in M_2. Intuitively, the best alternatives in the larger set M_1 are also the best in the smaller set M_2, or, to adapt an example of Amartya Sen himself [Sen, 1970, Sen, 2017]: the most subscribed to Youtubers in the world (M_1) that happen to be located in Europe (M_2) must also be among the most subscribed to Youtubers in Europe. Property C_4, sometimes called *property* β [Sen, 1969, Sen, 1970], says that if $M_2 \subseteq M_1$, then if there are alternatives chosen in M_2 that are also chosen in M_1, then any alternatives chosen in M_2 are also chosen in M_1. Intuitively, if some of the best alternatives in the smaller set M_2 also happen to be the best alternatives in the larger set M_1, then the best alternatives in M_2 are the among the best alternatives in M_1, or, using the Youtuber example: if x_1 and x_2 are the Youtubers in Europe (M_1) with the most subscribers (which implies that x_1 and x_2 have an equal number of subscribers), then x_1 is among the most subscried to Youtubers in the world (M_2) if and only if x_2 is as well.

Example 2.12: Choosing wisely

Consider the choice sets $M_1 = \{x_1, x_2, x_3\}$ and $M_2 = \{x_1, x_2\}$, and an agent whose choice function c is such that $c(M_1) = \{x_1\}$ and $c(M_2) = \{x_2\}$. Note that $M_2 \subseteq M_1$ but $c(M_1) \cap M_2 = \{x_1\}$ and $c(M_2) = \{x_2\}$, thereby contradicting property C_3. The agent is behaving strangely: when the menu consists of x_1, x_2 and x_3 it chooses x_1, signaling that it thinks x_1 is strictly better than x_2 and x_2, whereas when the menu is only x_1 and x_2, it chooses x_2, signaling that it thinks x_2 is better than x_1.

Consider, now, a different agent, whose choice function c' is such that $c'(M_1) = \{x_1\}$ and $c'(M_2) = \{x_1, x_2\}$. Property C_3 is satisfied, but property C_4 is not, since $x_2 \notin c'(M_1)$. This agent is also behaving strangely: when choosing among the elements of M_2 it signals that x_1 and x_2 are equally good, but when choosing among the elements of M_1 it signals that x_1 is strictly better than x_2.

We also mention the following property, expected to hold for any choice sets M_1 and M_2:

(C_5) If $M_2 \subseteq M_1$ and $c(M_1) \cap M_2 \neq \emptyset$, then $c(M_2) \subseteq c(M_1) \cap M_2$.

Property C_5 says that the elements chosen from the smaller set M_2 are also chosen from the larger set M_1, or, using the Youtubers example: if some of the most subscribed to Youtubers in the world (M_1) are located in Europe (M_2), then the most subscribed to Youtubers in Europe must also be among the most subscribed to Youtubers in the world. if it happens that some of the chosen elements in M_1 are also in M_2. As such, property C_5 is similar to C_4, but it is stronger than it: C_5 implies C_4, though it is not implied by it. However, properties C_3 and C_4 together imply C_5. We mention property C_5 mainly because of the symmetry it exhibits with C_3; it will prove more relevant in Chapter 3.

Example 2.12 already rationalizes a choosing agent's behavior in terms of alternatives that are good, or best in a certain range: this suggests that there exists a dimension

along which the agent ranks alternatives, and uses this ranking to guide its choices. This intuition is formalized through the instrument of a preference order $\leq$ on X, as described in Section 2.2, and expected to enjoy a two-way relation with a choice function c.

Firstly, a preference relation is used by the agent to determine its choices. Thus, given a preference order $\leq$ on X, the *$\leq$-induced choice function $c^{\leq}$ on X* is defined, for any choice set M, as:

$$c^{\leq}(M) \stackrel{\text{def}}{=} \min_{\leq} M.$$

Intuitively, given a menu M of alternatives in X, the choice over M consists of the best elements of M according to the preference relation $\leq$ on X.

Conversely, the agent's observed choice behavior can be used to construct, or reveal, the agent's preference relation on X. Thus, given a choice function c on X, the *c-revealed preference relation $\leq^c$ on X* is defined, for any $x_1, x_2 \in X$, as:

$$x_1 \leq^c x_2 \text{ if } x_1 \in c(\{x_1, x_2\}).$$

Intuitively, x_1 is considered at least as good as x_2 in $\leq^c$ if x_1 is chosen when the choice set contains exactly x_1 and x_2. If $\leq$ is a preference order on a set X and c is a choice function, then $\leq$ *represents* c (alternatively, c *is represented by* $\leq$), if, for any choice set $M \subseteq X$, it holds that:

$$c(M) = \min_{\leq} M.$$

In this section, we are interested in choice functions that can be represented by preference orders $\leq$ that are total preorders. The rationale for this is straightforward: the properties of a total preorder, though demanding and perhaps unrealistic, describe an agent who has complete information about the alternatives it is faced with, and can unfailingly identify the best alternatives out of any choice set. What properties does a choice function need to satisfy in order for it to be represented by a total preorder? The answer is provided by properties C_{1-4}.

Theorem 2.2 ([Sen, 1970])

If c is a choice function on X, then c satisfies properties C_{1-4} if and only if there exists a total preorder $\leq$ on X that represents c.

Intuitively, if c is a choice function satisfying properties C_{1-4} then the c-revealed preference relation $\leq^c$ is exactly the total preorder that satisfies the conditions in Theorem 2.2. What is more, the $\leq^c$-induced choice function on X is identical to c. Theorem 2.2 is similar, in many respects, to the representation results for belief change we will encounter in the following chapters.

Social choice

If $N = \{1, \ldots, n\}$ is a set of agents and X is a set of alternatives, a *$\mathcal{T}$-profile $\vec{\leq}$ on X* (alternatively, a *preference profile on X*), also written as $\vec{\leq} = (\leq_i)_{1 \leq i \leq n}$ is an n-tuple

$\vec{\leq} = (\leq_1, \ldots, \leq_n)$ of total preorders on X. Each total preorder $\leq_i$ in a preference profile $\vec{\leq}$ is assumed to correspond to an agent i in N. A *social choice function* sc is a function $sc\colon \mathcal{T}^n \to 2^X$, taking as input a preference profile $\vec{\leq}$ on X and returning a set $sc(\vec{\leq})$ of alternatives in X, called the *winning alternatives with respect to* $\vec{\leq}$ *and* X. A *social welfare function* sw is a function $sw\colon \mathcal{T}^n \to \mathcal{T}$, taking as input a $\mathcal{T}$-preference profile $\vec{\leq}$ on X and returning a total preorder $sw(\vec{\leq})$ on X.

The model for a social choice function is a voting rule, i.e., a function that processes preferences submitted by agents and outputs a set of winning candidated. The literature on social theory is, of course, rich in concrete proposals of voting rules and of ways to analyze them, and in the interest of brevity we defer to some standard sources for more in-depth material [Zwicker, 2016, Baumeister and Rothe, 2016]. Here we only introduce a few key notions that have been tweaked to fit in with the overall tenor of this work and will be referenced later. The first notion is the specific solution concept of a *weak Condorcet winner*. Thus, if X is a set of alternatives, $\vec{\leq} = (\leq_i)_{1\leq i\leq n}$ is a preference profile on X and x_1 and x_2 are alternatives in X, the *support* $\mathrm{supp}_X(x_1, x_2)$ *of* x_1 *over* x_2 *with respect to* $\vec{\leq}$ *and* X is defined as:

$$\mathrm{supp}_X(x_1, x_2) = \{i \in N \mid x_1 \leq_i x_2\},$$

i.e., the set of agents in N for whom x_1 is at least as good as x_2. Intuitively, we can think of x_1 and x_2 as matched up in a head to head election based on the preferences expressed in $\vec{\leq}$, and the support of x_1 over x_2 are the agents in N who see x_1 as at least as good as x_2. If x^* is an alternative in X, then x^* is a *weak Condorcet winner with respect to* $\vec{\leq}$ *and* X if:

$$|\mathrm{supp}_X(x^*, x)| \geq |\mathrm{supp}_X(x, x^*)|, \text{ for any } x \in X.$$

In other words, x^* is a weak Condorcet winner with respect to $\vec{\leq}$ and X if x^* is considered at least as good as any other alternative x by at least as many agents in N as those that consider x at least as good as x^*. Intuitively, an alternative x^* is a weak Condorcet winner with respect to $\vec{\leq}$ and X if x^* is in X and manages to defeat, or match, every other alternative in X in a head to head election.

The definition of a (weak) Condorcet winner is relativized to a set of alternatives X to allow for the possibility of varying X anywhere in between the universal set of alternatives and a subset of it. If agents are assumed to have preferences over the set of all possible alternatives, then we can restrict those preferences to a smaller set of alternatives and ask for the Condorcet winner relative to the restricted set. Clearly, a weak Condorcet winner relative to a set X of alternatives stays a weak Condorcet winner relative to a set $X' \subseteq X$ of alternatives. The converse, however, is not guaranteed to hold.

1	2	3	4	5
x_1	x_4	x_3	x_2	x_2
x_2	x_1	x_4	x_3	x_1
x_3	x_2	x_1	x_4	x_4
x_4	x_3	x_2	x_1	x_3

$\vec{\leq} = (\leq_i)_{1\leq i\leq 5}$ over $X = \{x_1, x_2, x_3, x_4\}$

1	2	3	4	5
x_1		x_3	x_2	x_2
x_2	x_1		x_3	x_1
x_3	x_2	x_1		
	x_3	x_2	x_1	x_3

$\vec{\leq}$ restricted to $X' = \{x_1, x_2, x_3\}$

Figure 2.3: On the left, the preference profile $\vec{\leq} = (\leq_i)_{1\leq i\leq 5}$ over the set of alternatives $X = \{x_1, x_2, x_3, x_4\}$; on the right, the same preference profile restricted to the set of alternatives $X' = \{x_1, x_2, x_3\}$. Higher, in this context, is better, such that for Doctor 1 alternative x_1 is the best, x_2 is the second best, and so on. We obtain that x_1 is a Condorcet winner with respect to $\vec{\leq}$ and X', but not with respect to $\vec{\leq}$ and X.

Example 2.13: Doctors in need of agreement

Recall the five doctors in Example 1.1 who have to agree on a common treatment for a novel respiratory disease, with the alternatives being a cocktail of drugs a and b, a alone, b alone or neither of these two drugs. If we denote these alternatives by x_1, x_2, x_3 and x_4, respectively, then each doctor can be thought of as having a preference order $\leq_i$, for $i \in \{1, 2, 3, 4, 5\}$, over the (universal) set of alternatives $X = \{x_1, x_2, x_3, x_4\}$, with these preferences depicted on the left in Figure 2.3. Figure 2.3 also depicts, on the right, the same preferences restricted to the set $X' = \{x_1, x_2, x_3\}$ of alternatives and depicted on the right in Figure 2.3.

Note that x_1 is the only weak Condorcet winner with respect to $\vec{\leq}$ and X'. Firstly, it holds that $\text{supp}_{X'}(x_1, x_2) = \{1, 2, 3\}$ and $\text{supp}_{X'}(x_2, x_1) = \{4, 5\}$, which means that (strictly) more agents (strictly) prefer x_1 to x_2. Secondly, it holds that $\text{supp}_{X'}(x_1, x_3) = \{1, 2, 5\}$ and $\text{supp}_{X'}(x_3, x_1) = \{3, 4\}$, which means that (strictly) more agents (strictly) prefer x_1 to x_3. Thus, x_1 conclusively defeats every other alternative in X' in a head to head election.

However, alternative x_1 ceases to be a weak Condorcet winner with respect to $\vec{\leq}$ and X: x_1 still defeats x_2 and x_3 in a head to head election, since the computations above are not changed; but x_1 is defeated by x_4, since $\text{supp}_X(x_1, x_4) = \{1, 5\}$ and $\text{supp}_X(x_4, x_1) = \{2, 3, 4\}$, i.e., more agents strictly prefer x_4 to x_1. What is more, since no alternative manages to defeat, or even match, all other alternatives in head to head elections, there is no weak Condorcet winner with respect to $\vec{\leq}$ and X.

We will use this relativized notion of a Condorcet winner in Section 5.2.

Since we will be concerned with belief merging operators that meet certain proportionality requirements, traditionally the preserve of multiwinner elections, we need a set of tools for thinking about proportionality in the context of social choice functions. Proportional

representation has been systematically studied in the social choice literature, notably in the case of Approval-Based Committee (ABC) elections [Faliszewski et al., 2017a]. An ABC election requires the set of alternatives X, a desired size of the committee k, and a particular type of preference profile.

In an ABC election voters are assumed to partition the set of alternatives into two sets: the alternatives they approve of, and the alternatives they do not approve of. In the context of the framework introduced here, such a preference order can be modeled by a preorder $\leq$ that consists of exactly two levels, i.e., there are two sets V and $X \setminus V$ such that $v_1 < v_2$ for any $v_1 \in V$ and $v_2 \in X \setminus V$ and $v \approx v'$ if $x, x' \in V$ or $v.v' \in X \setminus V$. If a preorder $\leq$ on X is of such a type, we call $\leq$ an *approval preference order*. If an agent i has an approval preference order $\leq_i$, then V_i is *agent i's approval ballot*. Since an approval ballot is just a set of alternatives from X, the set of all approval ballots is 2^X. An approval preference order is completely determined by the elements of V_i, so voters need only report their approval ballots for their preference to be completely specified, and we will base ABC elections on these ballots. Thus, an *approval profile* $\vec{V}$ is a tuple $\vec{V} = (V_1, \ldots, V_n)$, also written as $(V_i)_{1 \leq i \leq n}$, of approval ballots, with $(2^X)^n$ being the set of all approval profiles of length n.

If X is a finite set of alternative and n and k are integers, an *ABC social choice function* abc is a function $\text{abc}\colon (2^X)^n \to 2^X_k$, taking as input an approval profile and returning a set of alternatives, also called the *winning committees*, of size k. The ABC social choice function of immediate interest to us is called *Proportional Approval Voting* [Thiele, 1895]. It is based on the *harmonic function* h, which is a function $h\colon \mathbb{N} \to \mathbb{R}$, defined as:

$$h(\ell) = \sum_{i=1}^{\ell} \frac{1}{i},$$

with the added convention that $h(0) = 0$. Given an approval profile $\vec{V} = (V_i)_{1 \leq i \leq n}$ and a committee $W \subseteq X$ of size k, the PAV-*score of W with respect to $\vec{V}$* is defined as:

$$\text{PAV}(\vec{V}, W) = \sum_{i=1}^{n} h(|V_i \cap W|),$$

where h is the harmonic function. Given two committees W_1 and W_2 of size k, the PAV-induced preorder on 2^X_k, is defined as:

$$W_1 \geq^{\text{PAV}}_{\vec{V}} W_2 \text{ if } \text{PAV}(\vec{V}, W_1) \geq \text{PAV}(\vec{V}, W_2).$$

The PAV ABC function abc^{PAV} applied to the approval profile $\vec{V}$, for a desired size k of the committee, is defined as:

$$\text{abc}^{\text{PAV}}_k(\vec{V}) = \max_{\geq^{\text{PAV}}_{\vec{V}}}\{W \subseteq X \mid |W| = k\},$$

i.e., it outputs committees of size k that maximize the PAV score with respect to $\vec{V}$.

PAV	$x_1x_2x_3x_4$	$x_1x_2x_3x_4$	$x_1x_2x_3x_4$	$y_1y_2y_3y_4$	sum
$x_1x_2x_3x_4$	$h(4)$	$h(4)$	$h(4)$	$h(0)$	6.25
$x_1x_2x_3y_1$	$h(3)$	$h(3)$	$h(3)$	$h(1)$	**6.5**
$x_1x_2y_1y_2$	$h(2)$	$h(2)$	$h(2)$	$h(2)$	6
$x_1y_2y_2y_3$	$h(1)$	$h(1)$	$h(1)$	$h(3)$	4.83
$y_1y_2y_3y_4$	$h(0)$	$h(0)$	$h(0)$	$h(4)$	2.08
...	...	...	...	...	...

Table 2.2: PAV scores for a selection of committees of size 4. when the approval profile is $\vec{V} = (V_i)_{\leq i \leq 4}$. an optimal outcome according to the abc^{PAV} function is one that maximizes the PAV-score with respect to the profile $\vec{V}$.

Example 2.14: The PAV rule

Take a set of candidates $X \cup Y$, where $X = \{x_1, x_2, x_3, x_4\}$ and $Y = \{y_1, y_2, y_3, y_4\}$, and an approval profile $\vec{V} = (V_1, V_2, V_3, V_4)$ with $V_1 = V_2 = V_3 = \{x_1x_2x_3x_4\}$ and $V_4 = \{y_1y_2y_3y_4\}$. Suppose $k = 4$, i.e., the task is to choose committees of size 4. Intuitively, a proportional outcome would consist of three candidates from X and one from Y, to reflect the fact that supporters X outnumber supporters of Y in the profile $\vec{V}$ by a ratio of 3:1. Indeed, this is exactly the type of outcome the PAV rule will select. Table 2.2 depicts the PAV scores of a representative sample of possible winning committees. Note that an optimal outcome according to the abc^{PAV} function is one that maximizes the PAV-score with respect to the profile $\vec{V}$. In this case, this corresponds to committees consisting of three alternatives from X and one from Y, i.e., $\text{abc}_k^{\text{PAV}}(\vec{V}) = \{x_1x_2x_3y_1, x_1x_2x_3y_2, \dots\}$.

The PAV function is known to satisfy a number of desirable proportionality requirements [Aziz et al., 2017], and will serve as a template for proportional belief merging operators in Section 5.5.

CHAPTER 3

Varieties of Belief Change

In this chapter we introduce the main characters of our story: the established belief change operations of *revision*, *update* and *merging*, as well as the newer operation of *enforcement*, introduced by us in the build up to the present work. We rely mostly on existing work, with an eye towards how it connects to the choice material presented in Section 2.4. There will be a section on each of the aforementioned belief change operations, but we start by a brief detour in which we introduce belief change operators as a very abstract, very general notion.

An $\mathcal{L}^n$*-belief change operator* $\flat$ (alternatively, a *propositional belief change operator* $\flat$) is a function $\flat\colon \mathcal{L}^n \times \mathcal{L} \to \mathcal{L}$, taking as input a propositional profile $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$ and a propositional formula μ, and returning a propositional formula $\flat(\vec{\varphi}, \mu)$. If $\flat_1$ and $\flat_2$ are propositional belief change operators, then $\flat_1$ and $\flat_2$ are *equivalent* if, for any propositional profile $\vec{\varphi}$ and formula μ, it holds that $\flat_1(\vec{\varphi}, \mu) \equiv \flat_2(\vec{\varphi}, \mu)$.

We will typically *not* use either $\flat$ or $\flat(\vec{\varphi}, \mu)$ to refer to belief change operators or their output, instead opting for the more usual symbols: $\circ$ for revision, $\diamond$ for update, $\triangleright$ for enforcement and Δ for merging, with the output denoted as $\varphi \circ \mu$, $\varphi \diamond \mu$, $\varphi \triangleright \mu$ and $\Delta_\mu(\vec{\varphi})$, respectively. The aim of the present definition is to drive home the idea that the known belief change operators apply to the same types of objects and have the same type of output; indeed, that they are part of a single family. Despite the difference in terminology, a common methodology and related intuitions underlie all these operators, such that apprehension of one facilitates apprehension of the others. Thus, we will be most lavish with explanations when it comes to revision, with the understanding that most intuitions carry over to the other settings, and introduce extra motivation only when necessary.

At the most general level, the main difference between these operators lies in the nature of the input. Revision, update and enforcement are assumed to apply to a single agent and take only one propositional formula as input, i.e., $n = 1$, in which case we write

simply φ instead of (φ), and refer to φ as the agent's *prior* (or *initial*) *information*, to μ as the *new information*, and to the result as the *posterior information*. In the case of merging n is allowed to be any non-negative integer and here we make use of the full framework. Following convention, μ is referred to here as the *integrity constraint*.

A crucial notion in what is to come is that of an assignment on interpretations. If $\mathcal{L}_*$ is a set of formulas in $\mathcal{L}$ and $\mathcal{U}$ is the set of interpretations, an $\mathcal{L}_*$*-preference assignment* $\preccurlyeq$ *on* $\mathcal{U}$ (or, more simply, an $\mathcal{L}_*$*-assignment* $\preccurlyeq$ *on* $\mathcal{U}$) is a function $\preccurlyeq\colon \mathcal{L}_* \to 2^{\mathcal{U}\times\mathcal{U}}$, mapping every formula in $\mathcal{L}_*$ to a binary relation on $\mathcal{U}$ (i.e., a subset of $\mathcal{U}\times\mathcal{U}$). We typically write $\leq_\varphi$ instead of, as it were, $\preccurlyeq(\varphi)$, to denote the relation on $\mathcal{U}$ assigned to the formula φ by $\preccurlyeq$. We will use preference assignments to model preference relations on $\mathcal{U}$ that depend, in a way yet to be specified, on the formulas of $\mathcal{L}_*$, and we will see that belief change operators and assignments on intepretations are natural companions of each other. Each belief change operator will be characterized by its own type of assignment, and getting the right type of postulates to capture the right type of assignment will be one of the main goals of Chapter 6. Before we get there, however, we must start with the basics.

3.1 Revision

Perhaps the prototypical types of belief change is *revision*. Introduced as part of the landmark AGM paper on the logic of belief change [Alchourrón et al., 1985], revision quickly became the focus of attention for much subsequent work and, as any broad overview of belief change confirms [Gärdenfors, 1988, Hansson, 1999b, Peppas, 2008, Hansson, 2017, Fermé and Hansson, 2018], a *de facto* benchmark for testing new ideas and approaches to belief change. It is in this spirit that we present it here.

Revision models changes in prior information triggered by the availability of new, trusted information. In the most basic scenario the new information is accepted, and the agent in whose head this all happens modifies its existing beliefs accordingly. As such, revision is based on an intuition that runs through both commonsense reasoning as well as more sophisticated forms of inference such as Bayesian reasoning [Joyce, 2019, 3blue1brown, 2019]: that new evidence leads to new beliefs, but the new beliefs are not formed in a vacuum; rather, they are informed by prior beliefs. Where Bayesian models represent beliefs as probability distributions and changes in beliefs as changes in the corresponding probabilities, logical models of revision typically treat beliefs as sets of elements from a predetermined space of possibilities, and changes of beliefs as removals, or additions, to this set. This difference falls along the lines of the distinction drawn in epistemology between *belief* and *credence* [Schwitzgebel, 2019, Jackson, 2020]. According to this distinction, the term *belief* is reserved for a type of attitude that can span only three possible values: an agent either believes, disbelieves or withholds belief with respect to a statement; *credence*, on the other hand, is more a matter of degree, and is usually taken to indicate an agent's level of confidence with respect to a statement.

For our purposes, this distinction will be useful: beliefs, as the sort of tripartite attitude described above, can be modeled using propositional logic: an agent's belief in a statement

φ is encoded by the agent 'holding' φ, disbelief is the agent holding $\neg\varphi$, withholding belief is the agent holding neither. We will also make use of the more graded approach embodied by credences, but we will delegate that aspect to a different mechanism.

Even in the logic-based approach, there is significant variety in how beliefs are ultimately represented: the traditional AGM model uses propositional theories, i.e., sets of propositional formulas closed under consequence; other models rely on belief bases, i.e., sets of propositional formulas *not* required to be closed under the consequence relation [Hansson, 1999b]. Here we will follow the Katsuno-Mendelzon model [Katsuno and Mendelzon, 1992], where the agent's prior information, the newly acquired information and the result are all represented as single propositional formulas over a finite alphabet.

Example 3.1: Revision in the primatology community revisited

Humans use tools (a), chimpanzees are a different species from humans (b) and chimpanzees use tools (c): these are the facts placed under scrutiny by Jane Goodall's findings in the Gombe National Park in Tanzania, in the 1960s, as detailed in Example 1.2. In Example 2.1 it was mentioned that the prior beliefs of the primatology community are represented by the formula $\varphi = a \wedge b \wedge ((a \wedge b) \rightarrow \neg c)$, while Jane Goodall's findings can be boiled down in the propositional formula $\mu = c$. Note that $\varphi \models \neg c$ and, hence, $\varphi \wedge \mu$ is inconsistent: Jane Goodall's findings contradict the established consensus. There is no question that these findings are correct, but the original consensus cannot survive unchanged. What to do?

In this section we will focus on the Katsuno-Mendelzon model of revision [Katsuno and Mendelzon, 1992] and on the ways in which revision can be redescribed as a choice procedure. We will start by setting down, as logical postulates, some conditions a revision operator is expected to meet. Notably, the classical set of postulates that have been proposed turn out to define a class of operators that can be looked at in two ways: on the one hand as change, guided by logical postulates, of propositional theories in response to new data; and on the other hand as choice functions over outcomes that exploit plausibility rankings. This correspondence tells us that an agent faced with revision of its initial beliefs acts as if it chooses from a set of feasible outcomes the ones it considers most plausible.

Postulates

Formally, an *$\mathcal{L}$-revision operator* $\circ$ is a function $\circ\colon \mathcal{L} \times \mathcal{L} \rightarrow \mathcal{L}$, taking as input two propositional formulas, denoted here by φ and μ, and standing for the agent's prior information and the newly acquired information, respectively, and returning a propositional formula, denoted here by $\varphi \circ \mu$. Revision is a single-agent operation, i.e., φ and $\varphi \circ \mu$ are assumed to be located in one agent's head. The prior information φ is usually taken to encode *belief*, i.e., information the agent considers to be true, while μ is taken to encode information the agent learns must be true, and the examples of revision we will focus on

in this section follow this line. It should be mentioned, however, that there is nothing in the formalism tying it down to beliefs only. Thus, φ and μ can just as well represent things the agent would like to have versus things it is allowed to have, or actions the agent would like to do versus things it is allowed to do. The basic framework we will present here applies in the same manner to all these cases: the only condition is that the agent's attitude, be it a belief, a desire or an intention, be expressible as a propositional formula.

Since arbitrary belief change operators are neither interesting nor useful, rationality criteria are employed in order to narrow down the range of acceptable operators. Certain postulates have been a mainstay of the belief change literature since the AGM paper [Alchourrón et al., 1985], and are broadly accepted as minimal requirements for any belief revision operator. We will qualify this picture to some extent in this section, as well as in Chapter 4, but for now we proceed by going through the usual postulates. Though this is largely known material, we present it in a somewhat novel way: in order to highlight links to other belief change operators, we have adapted the names of the postulates to reflect a common naming scheme; we have also split the postulates into four disjoint groups, to reflect our understanding of how the postulates fit together; and we have slightly changed the order of presentation. We start, then, with the first group of postulates, applying to any propositional formulas φ, φ_1, φ_2, μ, μ_1 and μ_2:

(R_1) $\varphi \circ \mu \models \mu$.

(R_3) If μ is consistent, then $\varphi \circ \mu$ is consistent.

(R_4) If $\varphi_1 \equiv \varphi_2$ and $\mu_1 \equiv \mu_2$, then $\varphi_1 \circ \mu_1 \equiv \varphi_2 \circ \mu_2$.

One of the assumptions of revision is that new information originates with a trustworthy source, and therefore the agent should integrate it into its beliefs. This is exactly what postulate R_1, also known as the *Success* postulate [Fermé and Hansson, 2018], prescribes: revising prior information φ by newly acquired information μ involves a commitment to accept the newly acquired information, in the sense that $\varphi \circ \mu$ should imply μ. Postulate R_3, also known as the *Consistency* postulate [Fermé and Hansson, 2018], says that if the newly acquired information μ is consistent, then the posterior information should also be consistent. Postulate R_4, also known as the *Extensionality* postulate [Fermé and Hansson, 2018], says that the result depends only on the semantic content of the information involved (i.e., on the models of φ and μ) and is not sensitive to the way in which information is written down (i.e., on the syntax of φ and μ). This postulate can be thought of as preventing something like *framing effects* [Tversky and Kahneman, 1981, Kahneman, 2011], in which the same set of options is described in two different ways, and agents choose different things from it depending on the connotation, positive or negative, given to the options. We would like agents to make decisions, and, in particularly, to revise in ways that take into account only the information perceived as relevant: payoffs, in decision theory, and the semantic content of formulas in revision.

Though, of course, it could be argued that there are cases in which the syntax of formulas is relevant, there are also cases in which it is not and in standard revision models we want to focus on those cases.

Postulates R_1 and R_{3-4} are truly the theoretical minimum that any reasonable revision operator should satisfy, and for the rest of this work we will expect that any revision operator satisfies them.

Example 3.2: Different ways to revise a belief

Example 1.2 introduced us to a revision scenario, and Example 3.1 plaved it into a formal framework. Prior information is represented as $\varphi = a \wedge b \wedge ((a \wedge b) \rightarrow \neg c)$, saying that humans are the only species to use tools. New information is $\mu = c$, saying that chimpanzees use tools as well. Consider, now, revision operators $\circ^1$, $\circ^2$, $\circ^3$, $\circ^4$ and $\circ^5$, with $\varphi \circ^1 \mu = \neg a \wedge \neg b \wedge c$, $\varphi \circ^2 \mu = \neg a \wedge b \wedge c$, $\varphi \circ^3 \mu = a \wedge \neg b \wedge c$, $\varphi \circ^4 \mu = (\varphi \circ^2 \mu) \vee (\varphi \circ^3 \mu)$, and $\varphi \circ^5 \mu = a \wedge b \wedge c$. According to $\circ^1$, the result accepts μ and forgets all prior information, including the commonsense notions that humans use tools and chimpanzees are different form humans. According to $\circ^2$, the response to μ should be to accept the facts that chimpanzees are not human and that they use tools, but reject the notion that humans use tools—redefining, as it were, what a tool is. According to $\circ^3$, the right response is to accept that humans, as well as chimpanzees, use tools, but to reject the idea that chimpanzees are different from humans—redefining what it is to be human. The operator $\circ^4$ says that the answer is a tie between the answers given by $\circ^2$ and $\circ^3$: it could be either, or both. According to $\circ^5$, the right response is to accept that humans use tools, that chimpanzees are different from humans, and that chimpanzees use tools: what is rejected, in this case, is the belief that humans are the only species that uses tools (i.e., the implication $(a \wedge b) \rightarrow \neg c$ in φ). All five of these operators are consistent with postulates R_1 and R_{3-4}.

Recall Louis Leakey's telegram to Jane Goodall: "Now we must redefine tool, redefine Man, or accept chimpanzees as humans" (see Example 1.2). From our current lineup of operators, operators $\circ^2$ and $\circ^3$ seem closest to capturing Leakey's suggested possibilities, with the indecision between them captured by $\circ^4$. Operator $\circ^1$ spells out a very drastic revision policy: accept new information and forget the prior information: Operator $\circ^4$ stays close to the facts, and settles on an outcome that rejects the hypothesis that excluded one of them.

As Example 3.2 illustrates, postulates R_1 and R_{3-4} place very weak constraints on a revision operator: a new belief must be formed, based only on the semantic information provided, and it must imply μ. In particular, postulates R_1 and R_{3-4} say nothing about how a revision operator should behave across several instances of revision. For this latter task we have the following postulates, expected to hold for any propositional formulas φ, μ_1 and μ_2:

(R$_5$) $(\varphi \circ \mu_1) \wedge \mu_2 \models \varphi \circ (\mu_1 \wedge \mu_2)$.

(R$_6$) If $(\varphi \circ \mu_1) \wedge \mu_2$ is consistent, then $\varphi \circ (\mu_1 \wedge \mu_2) \models (\varphi \circ \mu_1) \wedge \mu_2$.

Postulates R$_{5-6}$, known as *Superexpansion* and *Subexpansion*, respectively [Fermé and Hansson, 2018], guide a belief change operator towards results that are coherent when the newly acquired information μ varies, according to a notion of coherence best illustrated by a series of examples.[1]

Example 3.3: A peculiar doctor

A doctor orders a test in order to diagnose a patient. On hearing that the test results are consistent with conditions a, b and c, the doctor says: "I think it's a". A second doctor points out that the test results are consistent only with conditions a and b, to which our doctor replies: "In that case I think it's b".

Framing this as a revision problem, suppose the doctor's prior information is φ (it does not matter, for this example, what φ actually is), and the two items of information it receives are $\mu_1 = a \vee b \vee c$ and $\mu_2 = (a \vee b) \wedge \neg c$, corresponding, respectively, to the facts that it could be conditions a, or b, or c, and that it could be a, or b, but not c. In response to μ_1, we can assume the doctor's response to be that the patient suffers from condition a alone, i.e., $\varphi \circ \mu_1 = a \wedge \neg b \wedge \neg c$. On hearing μ_2, the doctor changes its response to $\varphi \circ \mu_2 = \neg a \wedge b \wedge \neg c$. Note that $\mu_2 \models \mu_1$, and it holds that $(\varphi \circ \mu_1) \wedge \mu_2 \equiv a \wedge \neg b \wedge \neg c$, while $\varphi \circ (\mu_1 \wedge \mu_2) \equiv \varphi \circ \mu_2 \equiv \neg a \wedge b \wedge \neg c$. Consequently, neither of the postulates R$_5$ and R$_6$ is satisfied.

There is something odd about how the doctor in Example 3.3 jumps to conclusions: when told that μ_1 is the case, the doctor reasons as if they think that a is more likely than b, and when told that μ_2 is the case, the doctor reasons as if they think b is more likely than a. But μ_1 and μ_2, by themselves, have no bearing on which of the two is more likely! It seems, then, that either the doctor's estimates as to the relative likelihood of a and b are not constant over time, or, if they are constant, then in one case the doctor sets their mind on the less likely outcome. In the absence of any mitigating factors either of these behaviors seems irrational, and postulates R$_{5-6}$ are there to prevent it.

Similar demands as those of postulate R$_6$ are made by the following postulates, likewise expected to hold for any propositional formulas φ, μ, μ_1 and μ_2:

(R$_7$) If $\varphi \circ \mu_1 \models \mu_2$ and $\varphi \circ \mu_2 \models \mu_1$, then $\varphi \circ \mu_1 \equiv \varphi \circ \mu_2$.

(R$_8$) If $\mu \equiv \mu_1 \vee \mu_2$, then $(\varphi \circ \mu_1) \wedge (\varphi \circ \mu_2) \models \varphi \circ \mu$.

[1] The examples are inspired by examples from decision theory [Luce and Raiffa, 1957].

Postulate R_8 has been slightly re-phrased (normally there is no reference to μ and $\mu_1 \vee \mu_2$ is written everywhere instead of it), but at this point the difference from the usual statement is merely stylistic. This difference will only play a role in Chapter 6.

Postulate R_7 can be understood through the lens of Example 3.3: the eccentric revision behavior exhibited there constitutes a counterexample for R_7 too. For postulate R_8, consider the following example.

Example 3.4: Another peculiar doctor

A doctor orders two tests in order to diagnose a patient. The first test indicates that the patient might have one of conditions a or b, and the doctor says: "It must be a." The second test indicates that the patient might have one of conditions a or c, and, the doctor concludes, again: "It must be a". A third test indicates that the patient might have either of conditions a, b or c, to which the doctor's reaction is: "It must be b".

If $\mu_1 = (a \leftrightarrow \neg b)$ represents the result of the first test, $\mu_2 = (b \leftrightarrow \neg c)$ represents the result of the second test and $\mu_1 \vee \mu_2$ represents the result of the third test, then the doctor's reactions are given by $\varphi \circ \mu_1 \equiv \varphi \circ \mu_2 \equiv a \wedge \neg b \wedge \neg c$ and $\varphi \circ (\mu_1 \vee \mu_2) \equiv \neg a \wedge b \wedge \neg c$, which contradicts postulate R_8.

The behavior exhibited in Example 3.4 is just as odd as the one in Example 3.3: the doctor judges a to be more likely than b when only a and b are thought possible; then, a to be more likely than c when only a and c are thought possible; but b as more likely than a when all three options are on the table. Again, in the absence of any extra information about the possible outcomes (and for Example 3.4 we assume that the test results are the only information the doctor can rely on), this violates our intuitions about how likelihood assessments are supposed to work and postulate R_8 rules it out.

It becomes apparent that the role of postulates R_{5-8} is not to tell an agent *what* to believe in response to new information, but, rather, to ensure that the agent's protocol for changing its beliefs, whatever it is, is consistent when looked at on the whole, i.e., across varying pieces of incoming information. The natural way of rationalizing such revision behavior, as suggested by Examples 3.3 and 3.4, is by appeal to the agent's assessment of the likelihood of various outcomes with respect to each other, and postulates R_{5-8} seem to say something along the lines that this assessment should be constant throughout all instances of revision. This point of view will be made clearer when switching to the preference-based view of revision, but for now it must be mentioned that the intention is not to impose the entire set on a revision operator: indeed, as Proposition 3.1 shows, that would be redundant.

Proposition 3.1

If $\circ$ is a revision operator satisfying postulates R_1 and R_{3-6}, then $\circ$ also satisfies postulates R_7 and R_8.

Proof

The statement is trivially satisfied if either of μ_1 and μ_2 is inconsistent, so we will focus on the case when both are consistent.

For postulate R_7, note that the hypothesis implies that $\varphi \circ \mu_1 \equiv (\varphi \circ \mu_1) \wedge \mu_2$ and $\varphi \circ \mu_2 \equiv (\varphi \circ \mu_2) \wedge \mu_1$. By postulate R_3, all of these formulas are consistent, so we can use postulates R_{5-6} to obtain that:

$$\begin{aligned} \varphi \circ \mu_1 &\equiv (\varphi \circ \mu_1) \wedge \mu_2 && \text{(by hypothesis)} \\ &\equiv \varphi \circ (\mu_1 \wedge \mu_2) && \text{(by } R_{5-6}\text{)} \\ &\equiv (\varphi \circ \mu_2) \wedge \mu_1 && \text{(by } R_{5-6}\text{)} \\ &\equiv \varphi \circ \mu_2. && \text{(by hypothesis)} \end{aligned}$$

For postulate R_8, postulate R_1 implies that at least one of $(\varphi \circ (\mu_1 \vee \mu_2)) \wedge \mu_1$ and $(\varphi \circ (\mu_1 \vee \mu_2)) \wedge \mu_2$ is consistent. Without loss of generality, we can assume that $(\varphi \circ (\mu_1 \vee \mu_2)) \wedge \mu_1$ is consistent. This allows us to apply postulate R_6, which, together with postulate R_4, leads to:

$$\begin{aligned} \varphi \circ \mu_1 &\equiv (\varphi \circ (\mu_1 \vee \mu_2)) \wedge \mu_1 && \text{(by } R_4\text{)} \\ &\models \varphi \circ (\mu_1 \vee \mu_2) \wedge \mu_1. && \text{(by } R_6\text{)} \end{aligned}$$

This implies that $\varphi \circ \mu_1 \models \varphi \circ (\mu_1 \vee \mu_2)$ and, since $(\varphi \circ \mu_1) \wedge (\varphi \circ \mu_2) \models \varphi \circ \mu_1$, the conclusion follows immediately.

Proposition 3.1 can be extracted from the standard reference [Katsuno and Mendelzon, 1992], but here we present it in a form that makes it explicit that we do not need any additional postulates to derive it, whereas in the literature this is not always clear. Another way of stating the result is by saying that, in the presence of postulates R_1 and R_{3-5}, postulate R_6 implies postulates R_7 and R_8: the reason we would want to say this is because postulates R_7 and R_8 are usually meant to be thought of as alternatives to R_6, i.e., revision operators are standardly assumed to satisfy (besides postulates R_1 and R_{3-5}) either postulate R_6, or R_{7-8}. To give this a name, we say that a revision operator $\circ$ is *exhaustive* if it satisfies postulates R_1 and R_{3-6}, and *exclusive* if it satisfies postulates R_1, R_{3-5} and R_{7-8}. Proposition 3.1 shows that an exhaustive operator is also exclusive, i.e., if a revision operator satisfies postulate R_6 (besides postulates R_1 and R_{3-5}), it will also satisfy postulates R_{7-8} The converse, however, is not guaranteed.

Thus, postulates R_{7-8} can be thought of as weaker versions of R_6, describing revision

operators that satisfy less stringent demands. We will see shortly that the kind of demands postulate R_6 places on a revision operator is that it is able to compare any two outcomes, in a way that can be described as exhaustive (hence the name), whereas postulates R_{7-8} allow a revision operator to focus comparisons only on certain pairs of outcomes, i.e., it is more exclusive.

There is a conspicuous gap between postulates R_1 and R_3, suggesting that the picture is not yet complete. Indeed, this gap is usually occupied by the following postulate, expected to hold for any propositional formulas φ and μ:

(R_2) If $\varphi \wedge \mu$ is consistent, then $\varphi \circ \mu \equiv \varphi \wedge \mu$.

Postulate R_2, also known as the *Vacuity* postulate [Fermé and Hansson, 2018], says that if the newly acquired information μ does not contradict the prior information φ, the result is just the conjunction of μ and φ, i.e., if possible to simply add μ to φ, then do so. With this, postulate R_2 lays down what is usually considered the ideal, or uncontroversial, case for revision, in which nothing special needs to be done. Nonetheless, it is important to be aware that behind this innocent façade there lies a particular attitude towards the prior information in φ, i.e., that any parts of φ not explictly ruled out by the new information μ should be preserved. This attitude is rooted in what Peter Gärdenfors calls the principle of *informational economy*, guiding an agent to preserve as much of the information it has at its disposal as it can:

> ... information is in general not gratuitous, and unnecessary losses of information are therefore to be avoided. [Gärdenfors, 1988, p. 49]

In Gärdenfors' formulation, this principle takes on a normative value, i.e., it describes the way in which agents *should* approach their beliefs. But there is also some support for this view from the way in which humans actually treat their beliefs. Specifically, psychologist Harold Abelson has argued that beliefs are like possessions:

> One is reluctant to change any of one's major beliefs. They are familiar and comfortable, and a big change would upset the whole collection. [Abelson, 1986]

Postulate R_2 is usually assumed to be satisfied by any revision operator, and standard references include it in the list of core revision postulates [Katsuno and Mendelzon, 1992]. Since we will find reason to subject postulate R_2 to some scrutiny in Chapter 4, we set it apart from the other postulates here, and separate results that require it from those that do not.

Preferences over outcomes

Examples 3.3 and 3.4, and the discussion surrounding them, already foreshadow an important aspect of belief revision, namely that it feels natural to rationalize it in terms of assessments on the likelihood of different outcomes. This is a perspective we want to make precise now.

Talk of likelihoods suggests numerical measures, an angle that could be integrated into belief change by assigning probabilities to the items of belief being questioned. But we will refrain from doing so: the path that most models of logical belief change take is to rank items of belief only relative to each other, i.e., using ordinal rankings. In the Katsuno-Mendelzon model, which we take here as gospel, the entities ranked are interpretations, with every prior belief inducing such a ranking. Formally, this is instantiated by an *$\mathcal{L}$-assignment $\preccurlyeq$ on $\mathcal{U}$*, i.e., a function $\preccurlyeq\colon \mathcal{L} \to 2^{\mathcal{U}\times\mathcal{U}}$ mapping every propositional formula φ to a binary relation $\leq_\varphi$ on interpretations. The intention is that $\leq_\varphi$ stands for a ranking of interpretations in terms of their plausibility, with the assumption that the prior belief φ biases this ranking, i.e., we are talking about plausibility *given* φ. If w_1 and w_2 are interpretations, $w_1 \leq_\varphi w_2$ is to be read as saying that w_1 is considered at least as plausible as w_2 by an agent whose prior beliefs are φ.

As with revision operators, we want to make sure that assignments on interpretations are 'rational', i.e., that they possess certain desirable properties, and to clarify exactly in what way prior beliefs bias the relations that depend on them. We denote these properties by r_i, for $i \in \mathbb{N}$, and say that an $\mathcal{L}$-assignment $\preccurlyeq$ satisfies property r_i if $\leq_\varphi$ satisfies property r_i, for every propositional formula φ.

We start with the rationality properties, intended to hold for any propositional formula φ, and any interpretations w, w_1 and w_2:

(r_1) $w \leq_\varphi w$.

(r_2) If $w_1 \leq_\varphi w_2$ and $w_2 \leq_\varphi w_3$, then $w_1 \leq_\varphi w_3$.

(r_3) $w_1 \leq_\varphi w_2$ or $w_2 \leq_\varphi w_1$.

Property r_1 says that any interpretation w is at least as plausible as itself, i.e., that $\leq_\varphi$ is reflexive, for any propositional formula φ. Property r_2 says that $\leq_\varphi$ is transitive, for any propositional formula φ. In economics, where preference relations stand for actual preference, transitivity is a disputed property, with numerous examples showing that humans are never far from violating it [Luce, 1956, Quinn, 1990, Bar-Hillel and Margalit, 1988]. Nonetheless, there is some consensus around the idea that transitivity is useful if the agent is to be protected from types of malicious exploitation such as a *money pump* [Anand, 2009, Hansson and Grüne-Yanoff, 2018]: in a money pump scenario an agent would presumably be willing to swap an item in favor of a strictly better one for a non-negative fee; but if the agent's preferences contain a non-transitive cycle, then it will be coaxed into a series of exchanges that go on forever, or at least until the agent's purse

has run dry. It is not entirely obvious what a money-pump argument would look like in the context of a plausibility ranking on interpretations, but most accounts of belief revision assume some kind of transitive ranking notwithstanding, and we follow suit here. The rationale for having a transitive ranking will emerge soon, as we will see that in the presence of a non-transitive cycle revision under certain intuitive parameters ceases to be possible.

Together, properties r_1 and r_2 imply that $\leq_\varphi$ is a preorder on the set $\mathcal{U}$ of interpretations. Note that properties r_{1-2}, by themselves, do not guarantee that $\leq_\varphi$ is total: this is ensured by property r_3. We will call properties r_{1-3} the *structural properties* of an assignment, since they concern the properties of $\leq_\varphi$ solely as a ranking, regardless of the content of the prior belief φ. An $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations is *total* if $\preccurlyeq$ satisfies properties r_{1-3} and *partial* if $\preccurlyeq$ satisfies properties r_{1-2}.

The properties we consider next actually do factor in the role of φ, though the first one does so merely by identifying cases in which the content of φ is irrelevant. Thus, for any propositional formula φ and φ', the following property is expected to hold:

(r_4) If $\varphi \equiv \varphi'$, then it holds that if $w_1 \leq_\varphi w_2$, then $w_1 \leq_{\varphi'} w_2$.

Property r_4 implies that if φ and φ' are equivalent propositional formulas, then $\leq_\varphi = \leq_{\varphi'}$, i.e., the preorders that depend on them are identical. In other words, property r_4 makes sure that the preorder $\leq_\varphi$ depends only on the semantic content of φ, i.e., on its set of models, and not on the way φ is written. In other words, $\leq_\varphi$ is insensitive to the syntax of the formula encoding the prior information. An $\mathcal{L}$-assignment that satisfies property r_4 is called, correspondingly, *syntax insensitive*. There is an obvious parallel between property r_4 and postulate R_4, and this is intentional.

Properties r_{1-2} and r_4 are, in general, non-negotiable: we will assume that all $\mathcal{L}$-assignments on interpretations we deal with in this work are partial, at the very least, and insensitive to syntax.

The next set of properties establish the manner in which the prior belief φ is allowed to bias $\leq_\varphi$, and are expected to hold for any propositional formula φ and interpretations w_1 and w_2:

(r_5) If $w_1, w_2 \in [\varphi]$, then $w_1 \approx_\varphi w_2$.

(r_6) If $w_1, w_2 \in [\varphi]$, then $w_1 \not<_\varphi w_2$ and $w_2 \not<_\varphi w_1$.

(r_7) If $w_1 \in [\varphi]$ and $w_2 \notin [\varphi]$, then $w_1 <_\varphi w_2$.

As is apparent, properties r_{5-7} regulate the placement of the models of φ in the ranking $\leq_\varphi$. Properties r_{5-6} both say that models of φ are indistinguishable in terms of plausibility. According to property r_5, models of φ are considered equally plausible in $\leq_\varphi$, whereas

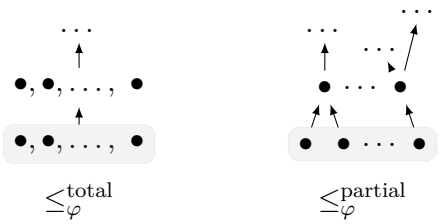


Figure 3.1: A schematic depiction of a total preorder $\leq_\varphi^{\text{total}}$ and a partial preorder $\leq_\varphi^{\text{partial}}$ in an r-faithful assignment. Bullets stand for interpretations. The situation where $w_1 \approx_\varphi^i w_2$ (depicted by bullets) is illustrated by drawing w_1 and w_2 on the same level and separating them by a comma, while the situation where neither $w_1 \leq_\varphi^i w_2$ nor $w_2 \leq_\varphi^i w_1$ is illustrated by drawing them apart. An arrow from interpretation w_1 to w_2 means that $w_1 <_\varphi^i w_2$, i.e., lower means strictly more plausible. Models of φ are shaded in light gray.

property r_6 requires that if w_1 and w_2 are models of φ, then one should not be considered strictly more plausible than the other. Separating these two intuitions might seem pedantic, but an understanding of the difference between the two will make life easier later on. The real difference between properties r_5 and r_6 kicks in if $\leq_{\vec{\varphi}}$ is partial: if $\leq_\varphi$ is total, then properties r_5 and r_6 coincide, in the sense that they are logically equivalent; but if $\leq_\varphi$ is partial, then they are different properties, with r_5 implying r_6 but not the other way around. Property r_7 says that models of φ are strictly more plausible than the other interpretations in the universe i.e., than non-models of φ.

An $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations is *r-faithful* if $\preccurlyeq$ satisfies properties r_6 and r_7. Note that, per the observation in the preceding paragraph, if $\preccurlyeq$ is total, then property r_6 implies r_5. Thus, a total r-faithful $\mathcal{L}$-assignment $\preccurlyeq$ actually satisfies property r_5 as well. A schematic illustration of preorders in a total and partial r-faithful assignment is given in Figure 3.1.

It is clear from the exposition above that the structural properties r_{1-3} are separate from properties r_{4-7}, in the sense that an assignment $\preccurlyeq$ can satisfy properties r_{1-3} without satisfying properties r_{4-7}. The two sets of properties also differ in their scope: properties r_{1-3} talk about the way in which $\preccurlyeq$ looks, whereas properies r_{4-7} about the influence of φ on $\leq_\varphi$ Nonetheless, it is a longstanding tradition in belief revision to focus on r-faithful $\mathcal{L}$-assignments that are also insensitive to syntax, or, as they are more commonly known, 'faithful assignments' [Katsuno and Mendelzon, 1992], since the properties we have introduced separately above are commonly bundled together into one package.[2] In r-faithful $\mathcal{L}$-assignments the relation $\leq_\varphi$ on interpretations is a preorder, either total or partial, in which models of φ are the $\leq_\varphi$-minimal, i.e., most plausible, outcomes. If insensitivity to syntax is added, as it usually is, then the preorder $\leq_\varphi$ depends only on the models of φ. In this section we will follow common practice in assuming that properties r_{5-7} are standard, and include them in the representation

[2]We add the 'r' qualifier in 'r-faithful' to distinguish such assignments from faithful assignments specific to other types of belief change, to come.

results, though gradually, so as to keep apart the intuitions around what depends on what. We will then subject properties r_{5-7} to more intense scrutiny in Chapter 4.

Revision as choice over outcomes

We have introduced two facets of an agent keen on revising its beliefs: on the one hand, a revision operator combines two propositional formulas into a new one, reflecting the change in belief; on the other hand, possible outcomes are ranked in terms of plausibility. We have hinted that the two facets are linked: here we finally show how they fit together.

The mechanism linking the two facets is that of choice: forming a new belief amounts to choosing, from a set of feasible outcomes, the most plausible ones. The feasible outcomes, in this case, are provided by the new information μ, and plausibility is provided by the preorder on outcomes: this is revision induced by the plausibility ranking. Conversely, inferring a plausibility ranking amounts to assuming, from an observed instance of revision, that the outcomes consistent with the result are considered more plausible than the outcomes that did not make the cut: this is plausibility revealed by revision behavior. The remainder of this section is devoted to spelling out the details of this picture.

The first direction involves using plausibility rankings on outcomes to determine how a belief is revised. Thus, given an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations, the *$\preccurlyeq$-induced revision operator* $\circ^{\preccurlyeq}$ is defined, for any propositional formulas φ and μ, by taking:

$$[\varphi \circ^{\preccurlyeq} \mu] \stackrel{\text{def}}{=} \min_{\leq_\varphi}[\mu].$$

The other direction involves reconstructing an agent's plausibility ranking over outcomes from its perceived revision behavior: we can tell what an agent thinks is more likely from the way it revises its beliefs. For this, recall that the $\mathcal{L}$-proxy $\varepsilon_{1,2}$ of two interpretations w_1 and w_2 is a propositional formula such that $[\varepsilon_{1,2}] = \{w_1, w_2\}$. We will distinguish between two ways of interpreting revision behavior: given a propositional belief change operator $\circ$ and a propositional formula φ, the *exhaustive $\circ$-revealed plausibility relation* $\leq_\varphi^{\text{exh}}$ and the *exclusive $\circ$-revealed plausibility relation* $\leq_\varphi^{\text{exc}}$ are defined, for any interpretations w_1 and w_2, respectively, as:

$$w_1 \leq_\varphi^{\text{exh}} w_2 \text{ if } w_1 \in [\varphi \circ \varepsilon_{1,2}],$$
$$w_1 \leq_\varphi^{\text{exc}} w_2 \text{ if } w_1 \in [\varphi \circ \varepsilon_{1,2}] \text{ and } w_2 \notin [\varphi \circ \varepsilon_{1,2}].$$

The *exhaustive revealed assignment* $\preccurlyeq^{\text{exh}}$ and *exclusive revealed assignment* $\preccurlyeq^{\text{exc}}$ are obtained by taking $\preccurlyeq^{\text{exh}}(\varphi) = \leq_\varphi^{\text{exh}}$ and $\preccurlyeq^{\text{exc}}(\varphi) = \leq_\varphi^{\text{exc}}$, for any propositional formula φ. The guiding intuition here is that if an agent leans toward outcome w_1 rather than outcome w_2 when it has the possibility of holding on to either of them, then this must be because the agent considers w_1 more plausible than w_2. That is, if $[\varphi \circ \varepsilon_{1,2}] = \{w_1\}$, then in both cases w_1 is considered better than w_2, i.e., $w_1 <_\varphi^{\text{exh}} w_2$ and $w_1 <_\varphi^{\text{exc}} w_2$. The difference between the two types of assignments lies in how they treat the case when both w_1 and w_2 are preserved, i.e., in the case when $[\varphi \circ \varepsilon_{1,2}] = \{w_1, w_2\}$. The exhaustive

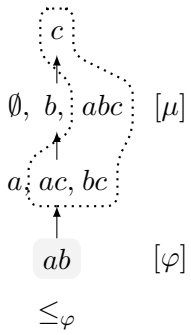


Figure 3.2: Total preorder $\leq_\varphi$, for $[\varphi] = \{ab\}$. Lower interpretations are better. The new information is $\mu = c$, and its models are surrounded by the dotted line. Revising φ by μ amounts to selecting the $\leq_\varphi$-minimal models of μ.

assignment assumes that the agent knows enough about the two outcomes (as it were, has exhaustive reasons) to conclude that they are equally likely, whereas the exclusive assignment merely infers that they cannot be compared, reserving judgment only for the exclusive case when w_1 is strictly better than w_2.

Revision by the $\mathcal{L}$-proxy of two interpretations w_1 and w_2 reveals a small piece of the agent's plausibility ranking over interpretations, namely the relative ranking of w_1 and w_2. Gluing these pieces together, one pair of interpretations at a time, yields the agent's full plausibility ranking $\leq_\varphi$. The endgame of this exercise is that we want the revealed plausibility relation to serve as a basis for explaining, or rationalizing, the revision behavior of the agent, in the same way that the preference relation of a single agent in Section 2.4 explains its choices across various menus. This is formalized by saying that if $\circ$ is an $\mathcal{L}$-revision operator and $\preccurlyeq$ is an $\mathcal{L}$-assignment on interpretations, then $\preccurlyeq$ *represents* $\circ$ (and $\circ$ *is represented by* $\preccurlyeq$) if, for any propositional formulas φ and μ, it holds that $[\varphi \circ \mu] = \min_{\leq_\varphi}[\mu]$.

Example 3.5: A monopoly on tool use after all?

We revisit our running revision scenario, detailed in Examples 1.2, 3.1 and 3.2 with Jane Goodall challenging the primatology status quo. This status quo is expressed as the propositional formula φ, with $[\varphi] = \{ab\}$, while Jane Goodall's challenge to it is the finding μ, with $[\mu] = \{c, ac, bc, abc\}$. Suppose that revision among the primatology community is guided by a total r-faithful assignment $\preccurlyeq$ that assigns to φ the total preorder $\leq_\varphi$ on interpretations in Figure 3.2. According to the preorder $\leq_{\vec{\varphi}}$, the state of the world ab is the most likely outcome: this corresponds with ab being the unique model of the prior belief φ, and is in agreement with properties r_{5-7}. Jane Goodall's finding reveals that the only viable outcomes are those consistent with μ i.e., the models of μ, and that a choice must be made as to which of these model will go into the new belief. How does the choice take place? We have that

$[\varphi \circ^{\preccurlyeq} \mu] = \min_{\leq_\varphi}[\mu] = \{ac, bc\}$, or, to put it differently, $\varphi \circ^{\preccurlyeq} \mu \equiv (a \wedge \neg b \wedge c) \vee (\neg a \wedge b \wedge c)$. That is, according to $\leq_\varphi$, the most plausible states of affairs consistent with μ, i.e., the $\leq_\varphi$-minimal models of μ, are ac and bc, and the rational course of action is to adopt them as the new belief.

Recall, as well, Louis Leakey's telegram to Jane Goodall from Example 1.2: "Now we must redefine tool, redefine Man, or accept chimpanzees as humans." Given how we model the problem, Leakey seems to think that the result of revising by μ should be either bc, i.e., chimpanzees use tools and are not human, but humans don't use tool (redefining tool), or ac, i.e., chimpanzees and humans use tools, but chimpanzees are human (accepting chimpanzees as humans). If we assume that Louis Leakey had the same prior belief φ as the rest of the primatology community but was revising according to a revision operator $\circ^{LL}$, then his telegram seems to suggest an inclination to conclude that ac or bc must be the case. In other words, narrowing down the choice to only these two outcomes (i.e., to $\varepsilon_{ac,bc} \equiv c \wedge (a \leftrightarrow \neg b)$), then Leakey seems to think that they cannot be distinguished, i.e., $[\varphi \circ^{LL} \varepsilon_{ac,bc}] = \{ac, bc\}$. In the exhaustive revealed ranking we would then infer that, for Leakey, it holds that $ac \approx^{\text{exh}}_{\varphi} bc$, and the preorder $\leq_\varphi$ in Figure 3.2 is consistent with this attitude, whereas in the exclusive revealed ranking we would infer that $ac \not\leq^{\text{exc}}_{\varphi} bc$ and $ac \not\leq^{\text{exc}}_{\varphi} bc$.

Example 3.5 illustrates that rankings of outcomes tell us something, not about the state of the world, but about what an agent thinks is more likely. They can be used to guide revision, or, conversely, they can be reconstructed, piece by piece, from putative revision behavior. But what ensures that a revision operator guided by a plausibility ranking on outcomes is rational? And how can we guarantee that the revealed rankings add up to a coherent whole? The answer depends entirely on the constraints imposed on the revision operator and on the plausibility rankings, with the representation results below showing that there is a close link between the revision postulates R_{1-8} and properties r_{1-7}. The first result shows that leaving out postulate R_2 and enforcing postulate R_6 results in revision policies that are represented by total assignments on interpretations that are also insensitive to syntax, which, we recall, means that $\leq_\varphi$ satisfies properties r_{1-4} (i.e., is a total preorder on $\mathcal{U}$), and that $[\varphi \circ \mu] = \min_{\leq_\varphi}[\mu]$, for any propositional formulas φ and μ. Recall, as well, that the $\mathcal{L}$-proxy of a set $\{w_1, \ldots, w_k\}$ of interpretations is a propositional formula $\varepsilon_{1,\ldots,k}$ such that $[\varepsilon_{1,\ldots,k}] = \{w_1, \ldots, w_k\}$.

Theorem 3.1

A revision operator $\circ$ satisfies postulates R_1 and R_{3-6} (i.e., is exhaustive) if and only if there exists an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that satisfies properties r_{1-4} (i.e., is total and insensitive to syntax) and represents the operator $\circ$.

Proof

("⇐") Assume, first, that we are given an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations such that, for any $\varphi \in \mathcal{L}$, $\leq_\varphi$ satisfies properties r_{1-4}. Since the $\preccurlyeq$-induced revision operator $\circ^\preccurlyeq$ is defined by taking $[\varphi \circ^\preccurlyeq \mu] \stackrel{\text{def}}{=} \min_{\leq_\varphi}[\mu]$, the proof amounts to showing that $\circ^\preccurlyeq$ satisfies postulates R_{1-4}.

Postulate R_1 follows from the fact that $\varphi \circ^\preccurlyeq \mu$ is a formula whose set of models is, by definition, a subset of $[\mu]$. Since $[\mu]$ is a finite set and, by properties r_{1-2}, $\leq_\varphi$ is a pre-order, we then have that $\min_{\leq_\varphi}[\mu] \neq \emptyset$, if $[\mu] \neq \emptyset$. This implies that postulate R_3 is satisfied. For postulate R_4 we have that if $\varphi_1 \equiv \varphi_2$, then, by property r_4, $\leq_{\varphi_1} = \leq_{\varphi_2}$. Clearly, then, if we also have that $\mu_1 \equiv \mu_2$, then it holds that $\min_{\leq_{\varphi_1}}[\mu_1] = \min_{\leq_{\varphi_2}}[\mu_2]$.

For postulate R_5, take $w_1 \in [(\varphi \circ^\preccurlyeq \mu_1) \wedge \mu_2]$: this means that $w_1 \in \min_{\leq_\varphi}[\mu_1] \cap [\mu_2]$, and we want to show that $w_1 \in \min_{\leq_\varphi}[\mu_1 \wedge \mu_2]$. Suppose, on the contrary, that $w_1 \notin \min_{\leq_\varphi}[\mu_1 \wedge \mu_2]$. Since we can derive, from our starting assumption, that $w_1 \in [\mu_1 \wedge \mu_2]$, it follows that $[\mu_1 \wedge \mu_2] \neq \emptyset$, and hence that $\min_{\leq_\varphi}[\mu_1 \wedge \mu_2] \neq \emptyset$. Thus there exists $w_2 \in \min_{\leq_\varphi}[\mu_1 \wedge \mu_2]$; since $w_1 \notin \min_{\leq_\varphi}[\mu_1 \wedge \mu_2]$ we then conclude that $w_2 <_\varphi w_1$. But w_1 and w_2 are both in $[\mu_1]$ and $w_1 \in \min_{\leq_\varphi}[\mu_1]$, which implies that $w_1 \leq_\varphi w_2$. We have arrived at a contradiction, and thus $w_1 \in \min_{\leq_\varphi}[\mu_1 \wedge \mu_2]$.

For postulate R_6, take $w_1 \in \min_{\leq_\varphi}[\varphi \circ^\preccurlyeq (\mu_1 \wedge \mu_2)]$. We want to show that $w_1 \in [(\varphi \circ^\preccurlyeq \mu_1) \wedge \mu_2]$. From the fact that $w_1 \in \min_{\leq_\varphi}[\varphi \circ^\preccurlyeq (\mu_1 \wedge \mu_2)]$ we infer that $w_1 \in [\mu_2]$, so all we have to show is that $w_1 \in [\varphi \circ^\preccurlyeq \mu_1]$. Suppose, on the contrary, that $w_1 \notin [\varphi \circ^\preccurlyeq \mu_1]$: we now use the assumption that $(\varphi \circ^\preccurlyeq \mu_1) \wedge \mu_2$ is consistent to conclude that there exists $w_2 \in [(\varphi \circ^\preccurlyeq \mu_1) \wedge \mu_2]$, which implies that $w_2 <_\varphi w_1$. But, since $w_1 \in \min_{\leq_\varphi}[\varphi \circ^\preccurlyeq (\mu_1 \wedge \mu_2)]$ and $w_2 \in [\mu_1 \wedge \mu_2]$, it also follows that $w_1 \leq_\varphi w_2$. This is a contradiction, and we conclude that $w_1 \in [\varphi \circ^\preccurlyeq \mu_1]$.

("⇒") Assume, now, that we are given an exhaustive belief change operator $\circ$, i.e., one that satisfies postulates R_1 and R_{3-6}. We will show that the exhaustive $\circ$-revealed assignment $\preccurlyeq^{\text{exh}}$ is the assignment we are looking for, i.e., $\leq_\varphi^{\text{exh}}$ satisfies properties r_{1-4}, for any propositional formula φ, and $[\varphi \circ \mu] = \min_{\leq_\varphi^{\text{exh}}}[\mu]$, for any propositional formula μ.

For property r_1 (reflexivity), take an interpretation w and the $\mathcal{L}$-proxy ε_w of w, i.e., a propositional formula such that $[\varepsilon_w] = \{w\}$. Notice that, using postulates R_1 and R_3, we can conclude that $[\varphi \circ \varepsilon_w] = \{w\}$. This implies that $w \leq_\varphi^{\text{exh}} w$.

For property r_2 (transitivity), assume there are interpretations w_1, w_2 and w_3 such that $w_1 \leq_\varphi^{\text{exh}} w_2$ and $w_2 \leq_\varphi^{\text{exh}} w_3$. We want to show that $w_1 \leq_\varphi^{\text{exh}} w_3$. We will do this in two steps. The first step consists in showing that $w_1 \in [\varphi \circ \varepsilon_{1,2,3}]$, where $\varepsilon_{1,2,3}$ is an $\mathcal{L}$-proxy of $\{w_1, w_2, w_3\}$, i.e., a propositional formula such that $[\varepsilon_{1,2,3}] = \{w_1, w_2, w_3\}$. First, notice that, by postulates R_1 and R_3, we have that $\emptyset \subset [\varphi \circ \varepsilon_{1,2,3}] \subseteq [\varepsilon_{1,2,3}]$. In

other words, $[\varphi \circ \varepsilon_{1,2,3}]$ contains at least one of the interpretations w_1, w_2 and w_3. We will do a case analysis to show that $w_1 \in [\varphi \circ \varepsilon_{1,2,3}]$.

Case 1. If $w_1 \in [\varphi \circ \varepsilon_{1,2,3}]$, the conclusion is immediate.

Case 2. If $w_2 \in [\varphi \circ \varepsilon_{1,2,3}]$, then $(\varphi \circ \varepsilon_{1,2,3}) \wedge \varepsilon_{1,2}$ is consistent. Using postulates R_{5-6} and R_4, and keeping in mind that $\varepsilon_{1,2,3} \wedge \varepsilon_{1,2} \equiv \varepsilon_{1,2}$, this implies that:

$$\begin{aligned}(\varphi \circ \varepsilon_{1,2,3}) \wedge \varepsilon_{1,2} &\equiv \varphi \circ (\varepsilon_{1,2,3} \wedge \varepsilon_{1,2}) && \text{(by } \mathrm{R}_{5-6}\text{)}\\ &\equiv \varphi \circ \varepsilon_{1,2}. && \text{(by } \mathrm{R}_4\text{)}\end{aligned}$$

By hypothesis, it holds that $w_1 \leq_\varphi^{\mathrm{exh}} w_2$, which, by the definition of the $\circ$-revealed exhaustive ranking, implies that $w_1 \in [\varphi \circ \varepsilon_{1,2}]$. Using this with the equivalence just derived, we arrive at the conclusion that $w_1 \in [\varphi \circ \varepsilon_{1,2,3}]$.

Case 3. If $w_3 \in [\varphi \circ \varepsilon_{1,2,3}]$, we infer that $(\varphi \circ \varepsilon_{1,2,3}) \wedge \varepsilon_{2,3}$ is consistent. Using, again, postulates R_{5-6} and R_4, and keeping in mind that $\varepsilon_{1,2,3} \wedge \varepsilon_{2,3} \equiv \varepsilon_{2,3}$, this implies that:

$$\begin{aligned}(\varphi \circ \varepsilon_{1,2,3}) \wedge \varepsilon_{2,3} &\equiv \varphi \circ (\varepsilon_{1,2,3} \wedge \varepsilon_{2,3}) && \text{(by } \mathrm{R}_{5-6}\text{)}\\ &\equiv \varphi \circ \varepsilon_{2,3}. && \text{(by } \mathrm{R}_4\text{)}\end{aligned}$$

Since $w_2 \in [\varphi \circ \varepsilon_{2,3}]$ (because $w_2 \leq_\varphi^{\mathrm{exh}} w_3$), we get that $w_2 \in [\varphi \circ \varepsilon_{1,2,3}]$. We can now reproduce the reasoning from Case 2 to conclude that $w_1 \in [\varphi \circ \varepsilon_{1,2,3}]$.

Rounding up the case analysis, we can conclude that $w_1 \in [\varphi \circ \varepsilon_{1,2,3}]$. With this in hand, we infer that $(\varphi \circ \varepsilon_{1,2,3}) \wedge \varepsilon_{1,3}$ is consistent, and we can apply the same blend of postulates R_{5-6} and R_4, keeping in mind that $\varepsilon_{1,2,3} \wedge \varepsilon_{1,3} \equiv \varepsilon_{1,3}$:

$$\begin{aligned}(\varphi \circ \varepsilon_{1,2,3}) \wedge \varepsilon_{1,3} &\equiv \varphi \circ (\varepsilon_{1,2,3} \wedge \varepsilon_{1,3}) && \text{(by } \mathrm{R}_{5-6}\text{)}\\ &\equiv \varphi \circ \varepsilon_{1,3}. && \text{(by } \mathrm{R}_4\text{)}\end{aligned}$$

Since $w_1 \in [\varphi \circ \varepsilon_{1,2,3}]$ and $w_1 \in [\varepsilon_{1,3}]$, we conclude that $w_1 \in [\varphi \circ \varepsilon_{1,3}]$. By the definition of $\leq_\varphi^{\mathrm{exh}}$, this implies that $w_1 \leq_\varphi^{\mathrm{exh}} w_3$.

Property r_3 (totality) follows from the fact for any two interpretations w_1 and w_2, there exists an $\mathcal{L}$-proxy $\varepsilon_{1,2}$ of $\{w_1, w_2\}$, and postulate R_3 guarantees that at least one of w_1 and w_2 is in $[\varphi \circ \varepsilon_{1,2}]$. Property r_4 follows by using postulate R_4, i.e., the fact that the definition of $\leq_\varphi^{\mathrm{exh}}$ is not sensitive in any way to the syntax of φ.

The last thing we have to show is that the exhaustive $\circ$-revealed assignments represents $\circ$, i.e., that $[\varphi \circ \mu] = \min_{\leq_\varphi^{\mathrm{exh}}}[\mu]$, for any propositional formula μ. We do this by showing the double inclusion.

("$\subseteq$") Take, first, $w_1 \in [\varphi \circ \mu]$, and some arbitrary interpretation $w_2 \in [\mu]$. Applying postulates R_5 and R_4 and keeping in mind that, because $[\varepsilon_{1,2}] \subseteq [\mu]$, it holds that

$\mu \wedge \varepsilon_{1,2} \equiv \varepsilon_{1,2}$, we have that:

$$\begin{aligned} (\varphi \circ \mu) \wedge \varepsilon_{1,2} &\models \varphi \circ (\mu \wedge \varepsilon_{1,2}) && \text{(by } \mathrm{R}_5\text{)} \\ &\equiv \varphi \circ \varepsilon_{1,2}. && \text{(by } \mathrm{R}_4\text{)} \end{aligned}$$

Since $w_1 \in [(\varphi \circ \mu) \wedge \varepsilon_{1,2}]$, it follows that $w_1 \in [\varphi \circ (\mu \wedge \varepsilon_{1,2})]$ and then that $w_1 \in [\varphi \circ \varepsilon_{1,2}]$. Thus, $w_1 \leq_\varphi^{\text{exh}} w_2$ and, keeping in mind that w_2 was arbitrarily chosen among the models of μ, we obtain that $w_1 \in \min_{\leq_\varphi^{\text{exh}}}[\mu]$.

("$\supseteq$") Take, now, $w_1 \in \min_{\leq_\varphi^{\text{exh}}}[\mu]$. We want to show that $w_1 \in [\varphi \circ \mu]$. Suppose, on the contrary, that $w_1 \notin [\varphi \circ \mu]$. Since, due to our assumption, it follows that μ is consistent, we have, by postulate R_3, that there exists $w_2 \in [\varphi \circ \mu]$. Using postulates R_4 and R_6, we have that:

$$\begin{aligned} \varphi \circ \varepsilon_{1,2} &\equiv \varphi \circ (\mu \wedge \varepsilon_{1,2}) && \text{(by } \mathrm{R}_4\text{)} \\ &\models (\varphi \circ \mu) \wedge \varepsilon_{1,2}. && \text{(by } \mathrm{R}_6\text{)} \end{aligned}$$

By assumption, we have that $w_1 \notin [\varphi \circ \mu]$, and from this it follows that $w_1 \notin [(\varphi \circ \mu) \wedge \varepsilon_{1,2}]$ and, using the implications just derived, it holds that $w_1 \notin [\varphi \circ \varepsilon_{1,2}]$, and hence $w_2 <_\varphi^{\text{exh}} w_1$. But we also have that $w_1 \in \min_{\leq_\varphi^{\text{exh}}}[\mu]$ and $w_2 \in [\mu]$, which implies that $w_1 \leq_\varphi^{\text{exh}} w_2$. We have thus arrived at a contradiction.

The result of Theorem 3.1 is not exactly new, and can be easily extracted from the literature [Katsuno and Mendelzon, 1992]; it is certainly present in Hans Rott's work [Rott, 2001]. Nonetheless, in many standard presentations postulates R_1 and R_{3-6} are taken together with postulate R_2, creating the impression that the postulates are inextricably tied together. Our purpose here is to separate the postulates that guarantee the structural properties r_{1-3} of the assignment $\preccurlyeq$, from the postulates that say where the models of φ should be placed in that preorder. Theorem 3.1 allows us to see that these are two distinct issues.

The second result shows that leaving out postulate R_2 and enforcing postulates R_{7-8} instead of R_6 results in revision policies that are represented by partial assignments on interpretations that are also insensitive to syntax. Recall, this means that $\leq_\varphi$ satisfies properties r_{1-2} (i.e., is a partial preorder on $\mathcal{U}$) and r_4, and that $[\varphi \circ \mu] = \min_{\leq_\varphi}[\mu]$, for any propositional formulas φ and μ.

Theorem 3.2

A revision operator $\circ$ satisfies postulates R_1, R_{3-5} and R_{7-8} (i.e., is exclusive) if and only if there exists an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that satisfies properties r_{1-2} and r_4 (i.e., is partial and insensitive to syntax) and represents the operator $\circ$.

Proof

("⇐") Starting from an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that satisfies properties r_{1-2} and r_4, the argument that $\circ^{\preccurlyeq}$ satisfies postulates R_1 and R_{3-5} is entirely similar to the argument for Theorem 3.1.

For postulate R_7, we have to show that $\min_{\leq_\varphi}[\mu_1] = \min_{\leq_\varphi}[\mu_2]$. Take, then, $w_1 \in \min_{\leq_\varphi}[\mu_1]$ and suppose that $w_1 \notin \min_{\leq_\varphi}[\mu_2]$. By postulate R_3 we have that $\varphi \circ^{\preccurlyeq} \mu_2$ is consistent, which implies that there exists $w_2 \in \min_{\leq_\varphi}[\mu_2]$ such that $w_2 <_\varphi w_1$. By the assumption of R_7, it holds that $\varphi \circ^{\preccurlyeq} \mu_2 \models \mu_1$, from which it follows that $w_2 \in [\mu_1]$ and, since $w_1 \in \min_{\leq_\varphi}[\mu_1]$, it follows that $w_2 \not<_\varphi w_1$, We have arrived, in this way, at a contradiction, and thus we conclude that $\varphi \circ^{\preccurlyeq} \mu_1 \models \varphi \circ^{\preccurlyeq} \mu_2$. The argument that $\varphi \circ^{\preccurlyeq} \mu_2 \models \varphi \circ^{\preccurlyeq} \mu_1$ is entirely similar. Together, these two facts imply that $\varphi \circ^{\preccurlyeq} \mu_1 \equiv \varphi \circ^{\preccurlyeq} \mu_2$.

For postulate R_8, take $w \in \min_{\leq_\varphi}[\mu_1] \cap \min_{\leq_\varphi}[\mu_2]$ and suppose that $w \notin \min_{\leq_\varphi}[\mu_1 \vee \mu_2]$. By postulate R_3 we have that $\min_{\leq_\varphi}[\mu_1 \vee \mu_2] \neq \emptyset$, and thus there exists $w' \in \min_{\leq_\varphi}[\mu_1 \vee \mu_2]$ such that $w' <_\varphi w$. Since, by postulate R_1, it holds that $w' \in [\mu_1 \vee \mu_2]$, we conclude that w' has to be an element of $[\mu_1]$ or of $[\mu_2]$. This leads to a contradiction with the fact that $w \in \min_{\leq_\varphi}[\mu_1] \cap \min_{\leq_\varphi}[\mu_2]$, because from this we are forced to conclude that $w' \not<_\varphi w$.

("⇒") Given a revision operator $\circ$ that satisfies postulates R_1, R_{3-5} and R_{7-8}, we will show that the exclusive $\circ$-induced $\mathcal{L}$-assignment $\preccurlyeq^{\text{exc}}$ on interpretations is the assignment we are looking for, i.e., $\leq_\varphi^{\text{exc}}$ satisfies properties r_{1-2} and r_4, for any propositional formula φ and, for any propositional formula μ, it holds that $[\varphi \circ \mu] = \min_{\leq_\varphi^{\text{exc}}}[\mu]$.

For r_1 (reflexivity), note that, by postulates R_1 and R_3, it holds, for any interpretation w, that $[\varphi \circ \varepsilon_w] = \{w\}$. This implies that $w \leq_\varphi^{\text{exc}} w$.

To show that $\leq^{\text{exc}}$ satisfies property r_3 (i.e., is transitive), take interpretations w_1, w_2 and w_3 such that w_1, w_2 and w_3 are pairwise distinct and $w_1 \leq_\varphi^{\text{exc}} w_2$ and $w_2 \leq_\varphi^{\text{exc}} w_3$. We show, first, that $[\varphi \circ \varepsilon_{1,2,3}] = \{w_1\}$, where $\varepsilon_{1,2,3}$ is an $\mathcal{L}$-proxy of the set $\{w_1, w_2, w_3\}$, i.e., a propositional formula such that $[\varepsilon_{1,2,3}] = \{w_1, w_2, w_3\}$. To that end, note that, by postulates R_1 and R_3, we have that $\emptyset \subset [\varphi \circ \varepsilon_{1,2,3}] \subseteq [\varepsilon_{1,2,3}]$. Suppose, now, that $w_2 \in [\varphi \circ \varepsilon_{1,2,3}]$. This means that $w_2 \in [(\varphi \circ \varepsilon_{1,2,3}) \wedge \varepsilon_{1,2}]$. Applying postulates R_5 and R_4 we obtain that:

$$\begin{aligned} (\varphi \circ \varepsilon_{1,2,3}) \wedge \varepsilon_{1,2} &\models \varphi \circ (\varepsilon_{1,2,3} \wedge \varepsilon_{1,2}) && \text{(by } R_5\text{)} \\ &\equiv \varphi \circ \varepsilon_{1,2}. && \text{(by } R_4\text{)} \end{aligned}$$

But this implies that $w_2 \in [\varphi \circ \varepsilon_{1,2}]$, which is a contradiction, since, by hopthesis we have that $w_1 \leq_\varphi^{\text{exc}} w_2$, which, by the definition of $\leq_\varphi^{\text{exc}}$, implies that $[\varphi \circ \varepsilon_{1,2}] = \{w_1\}$. This shows that $w_2 \notin [\varphi \circ \varepsilon_{1,2,3}]$. Applying the same strategy to $[(\varphi \circ \varepsilon_{1,2,3}) \wedge \varepsilon_{2,3}]$, it

follows that $w_3 \notin [\varphi \circ \varepsilon_{1,2,3}]$. Thus, the only remaining possibility is that $[\varphi \circ \varepsilon_{1,2,3}] = \{w_1\}$.

With this result in hand, we have that $\varphi \circ \varepsilon_{1,2,3} \models \varepsilon_{1,3}$. It is straightforward to see that $\varphi \circ \varepsilon_{1,3} \models \varepsilon_{1,2,3}$, which allows us to apply postulate R_7 and infer that $\varphi \circ \varepsilon_{1,2,3} \equiv \varphi \circ \varepsilon_{1,3}$. This means that $[\varphi \circ \varepsilon_{1,3}] = \{w_1\}$ and, by the definition of $\leq_\varphi^{\text{exc}}$, it follows that $w_1 \leq_\varphi^{\text{exc}} w_3$.

For r_4 (i.e., insensitivity to the syntax of φ and μ), the argument is entirely similar to the one given in the proof of Theorem 3.1.

Finally, we show that $[\varphi \circ \mu] = \min_{\leq_\varphi^{\text{exc}}}[\mu]$ by double inclusion.

("$\subseteq$") Take, first, $w \in [\varphi \circ \mu]$ and suppose $w \notin \min_{\leq_\varphi^{\text{exc}}}[\mu]$. This means that there exists $w' \in \min_{\leq_\varphi^{\text{exc}}}[\mu]$ such that $w' <_\varphi^{\text{exc}} w$, which in turn implies that $[\varphi \circ \varepsilon_{w,w'}] = \{w'\}$. But, by postulates R_5 and R_4, we have that:

$$\begin{aligned} (\varphi \circ \mu) \wedge \varepsilon_{w,w'} &\models \varphi \circ (\mu \wedge \varepsilon_{w,w'}) && \text{(by } R_5\text{)} \\ &\equiv \varphi \circ \varepsilon_{w,w'}. && \text{(by } R_4\text{)} \end{aligned}$$

Thus, since $w \notin [\varphi \circ \varepsilon_{w,w'}]$ but $w \in [\varepsilon_{w,w'}]$, it follows that $w \notin [\varphi \circ \mu]$, which is a contradiction.

("$\supseteq$") Take, now, $w \in \min_{\leq_\varphi^{\text{exc}}}[\mu]$ and suppose $w \notin [\varphi \circ \mu]$, and an arbitrary $w_i \in [\mu]$. Since $w \in \min_{\leq_\varphi^{\text{exc}}}[\mu]$, it cannot be the case that $w_i <_\varphi^{\text{exc}} w$, which implies that $w \in [\varphi \circ \varepsilon_{w,w_i}]$: to see why, suppose that $w \notin [\varphi \circ \varepsilon_{w,w_i}]$; by postulates R_1 and R_3, $[\varphi \circ \varepsilon_{w,w_i}]$ needs to be a non-empty subset of $[\varepsilon_{w,w_i}]$, and this implies that $[\varphi \circ \varepsilon_{w,w_i}] = \{w_i\}$, hence $w_i <_\varphi^{\text{exc}} w$. Applying postulate R_8 for every $w_i \in [\mu]$, and keeping in mind that $\bigvee_{w_i \in [\mu]} \varepsilon_{w,w_i} \equiv \mu$, we obtain that:

$$\begin{aligned} \bigwedge_{w_i \in [\mu]} (\varphi \circ \varepsilon_{w,w_i}) &\models \varphi \circ \Big(\bigvee_{w_i \in [\mu]} \varepsilon_{w,w_i} \Big) && \text{(by } R_5\text{)} \\ &\equiv \varphi \circ \mu. && \text{(by } R_4\text{)} \end{aligned}$$

It then follows that $w \in [\varphi \circ \mu]$.

Theorems 3.1 and 3.2 make it official: the behavior of an agent revising its beliefs according to postulates R_1, R_{3-5} and either postulate R_6 or postulates R_{7-8}, can be rationalized using preorders on outcomes, such that the outcomes the agent ends up accepting as part of its revised belief are the most plausible outcomes consistent with the new information. It is as if the new information provides a menu of allowable alternatives, and the agent chooses the best outcomes from this menu to believe.

Indeed, the similarity of this perspective with the rational choice framework for a single

agent presented in Section 2.4 runs deeper, as an $\mathcal{L}$-revision operator $\circ$ can be seen to be a choice function over the set $\mathcal{U}$ of interpretations, with $[\mu]$ as the choice set. Contemplation of postulates R_1, R_3 and R_{5-6} quickly reveals the parallel to the axioms for choice functions: viewed semantically, postulates R_1 and R_3 say that $[\varphi \circ \mu] \subseteq [\mu]$ and that, if $[\mu] \neq \emptyset$, then $[\varphi \circ \mu] \neq \emptyset$, which coincides with properties C_1 and C_2, respectively, of a choice function. Since these properties can be taken to be constitutive of a choice function, we could even prove a mini-result saying that any revision operator satisfying postulates R_1 and R_{3-4} is equivalent to a choice function on the set $\mathcal{U}$ of interpretations satisfying properties C_{1-2}: this is sufficiently obvious, however, to leave it as an observation.

Moving further, it can be seen that postulates R_5 and R_6 coincide with properties C_3 and C_5, respectively. Correspondingly, deviations from them are similar in spirit: Examples 3.3 and 3.4, showing agents that revise in ways inconsistent with postulates R_{5-8} are, on a close reading, entirely consonant with Example 2.12, showing an agent that chooses in ways inconsistent with properties C_{3-4}. The parallel is entirely justified, as both types of agents exhibit the same kind of pathological behavior when choosing among a set menu: the doctors in Examples 3.3 and 3.4 are just choosing odd things to believe, or, to be more precise, they revise in ways that are not immediately rationalizable by unique plausibility relations on outcomes. In the same spirit, Theorem 3.1 can be seen as a direct analogue of Theorem 2.2. Postulate R_4 has no analogue in the choice framework, since no distinction is made there between syntax and semantics.

We can see unfolding here a point that has been made before in belief change [Rott, 1992, Schulte, 1999, Rott, 2001, Bonanno, 2009, Arló-Costa and Pedersen, 2010], namely that the choice perspective is integral to the workings of a revision operator. And, while we will want to take up this point and explore it further, it is important to not be too carried away by its significance. We certainly do not want to suggest that either properties C_{1-5} or postulates R_1 and R_{3-8} uniquely characterize rational choice or rational belief change, since examples to the contrary are readily available [Sen, 1977, Olsson, 2003, Kahneman, 2011]: rationality, as we have mentioned before, comes in many flavors, and what is rational in one type of situation may not be rational in another. We can thus presume that the properties covered so far in Section 2.4 and in the present section touch on only a very small part of the whole gamut of rational behavior, and, inde. So, while, we do not want to give this particular formulation undue weight, we do want to use the broader implication, i.e., that belief change is a type of change, to explore the variety of change procedures that could count as rational.

Along these lines, we can start thinking of the role of postulate R_2 in the lineup of desirable revision postulates. One thing that emerges from Theorems 3.1 and 3.2 is that postulates R_1 and R_{3-5}, together with either postulate R_6 or postulates R_{7-8}, regulate only the structural properties of the preorder $\leq_\varphi$ in an assignment, and say nothing about how the prior information φ biases $\leq_\varphi$, i.e., about the position of the models of φ in $\leq_\varphi$. This latter aspect, as we will see in Theorems 3.3 and 3.4, is traditionally enforced through postulate R_2. For these results keep in mind that an $\mathcal{L}$-proxy of a pair $\{w_1, w_2\}$

of interpretations is a propositional formula $\varepsilon_{1,2}$ such that $[\varepsilon_{1,2}] = \{w_1, w_2\}$, and that the lessons of Theorems 3.1 and 3.2 are that exhaustive and exclusive $\mathcal{L}$-revision operators are guaranteed to be represented by total and partial $\mathcal{L}$-assignments on interpretations, respectively.

Theorem 3.3

If a revision operator $\circ$ satisfies postulates R_1 and R_{3-6} (i.e., is exhaustive) and $\preccurlyeq$ is an $\mathcal{L}$-assignment on interpretations that satisfies properties r_{1-4} (i.e., is total and syntax insensitive) and represents the operator $\circ$, then $\circ$ satisfies postulate R_2 if and only if $\preccurlyeq$ satisfies properties r_{5-7} (i.e., is r-faithful).

Proof

("$\Rightarrow$") We start with a revision operator $\circ$ satisfies postulates R_1 and R_{3-6} and a total, syntax insensitive $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that represents it. Consider, now, a propositional formula φ and two interpretations w_1 and w_2 such that w_1 and w_2 are models of φ. Using postulate R_2, we can conclude that $[\varphi \circ \varepsilon_{1,2}] = \{w_1, w_2\}$, which, together with the fact that $\leq_\varphi$ is total, implies that $w_1 \approx_\varphi w_2$, showing that property r_5 is satisfied. If $w_1 \in [\varphi]$ and $w_2 \notin [\varphi]$, then with postulate R_2 again we conclude that $[\varphi \circ \varepsilon_{1,2}] = \{w_1\}$, which implies that $w_1 <_\varphi w_2$, showing that property r_7 is satisfied.

("$\Leftarrow$") Conversely, we have to show that if $[\varphi \wedge \mu] \neq \emptyset$, then $\min_{\leq_\varphi}[\mu] = [\varphi \wedge \mu]$. We can do this by showing the double inclusion.

("$\subseteq$") Take $w_1 \in \min_{\leq_\varphi}[\mu]$ and suppose $w_1 \notin [\varphi \wedge \mu]$. Since $w_1 \in [\mu]$, by postulate R_1, the latter fact implies that $w_1 \notin [\varphi]$. Since $[\varphi \wedge \mu] \neq \emptyset$, there exists an interpretation $w_2 \in [\varphi \wedge \mu]$. We infer from this that $w_2 \in [\varphi]$ and, together with property r_7, that $w_2 <_\varphi w_1$. But $w_1 \in \min_{\leq_\varphi}[\mu]$, so this creates a contradiction.

("$\supseteq$") Take $w_1 \in [\varphi \wedge \mu]$ and an arbitrary interpretation $w_2 \in [\mu]$. Using properties r_5 and r_7, and keeping in mind that $\leq_\varphi$ is total, we conclude that $w_1 \leq_\varphi w_2$, which implies that $w_1 \in \min_{\leq_\varphi}[\mu]$.

Theorem 3.3 takes care of the case when $\circ$ satisfies the stronger postulate R_6 and is represented by a total assignment. We can obtain a similar result for the case when $\circ$ satisfies the weaker postulates R_{7-8} instead of R_6, and is represented by a partial assignment.

Theorem 3.4

If a revision operator $\circ$ satisfies postulates R_1, R_{3-5} and R_{7-8} (i.e., is exclusive) and $\preccurlyeq$ is an $\mathcal{L}$-assignment on interpretations that satisfies properties r_{1-2} and r_4

(i.e., is partial and syntax insensitive) and represents the operator $\circ$, then $\circ$ satisfies postulate R_2 if and only if $\preccurlyeq$ satisfies properties r_6 and r_7 (i.e., is r-faithful).

Proof

The proof here follows the same lines as the proof for Theorem 3.3, so more intuitions can be gleaned from there.

("$\Rightarrow$") If $w_1, w_2 \in [\varphi]$, then using postulate R_2 gives us that $[\varphi \circ \varepsilon_{1,2}] = \{w_1, w_2\}$. However, since $\leq_\varphi$ is not guaranteed to be total, we cannot conclude that $w_1 \approx_\varphi w_2$; however, we can conclude that $w_1 \not<_\varphi w_2$ and $w_2 \not<_\varphi w_1$, which means that property r_6 is satisfied. If $w_1 \in [\varphi]$ and $w_2 \notin [\varphi]$, we obtain using postulate R_2 that $[\varphi \circ \varepsilon_{1,2}] = \{w_1\}$, which implies that $w_1 <_\varphi w_2$, showing that property r_7 is satisfied.

("$\Leftarrow$") We have to show that if $[\varphi \wedge \mu] \neq \emptyset$, then $\min_{\leq_\varphi}[\mu] = [\varphi \wedge \mu]$. The proof that $\min_{\leq_\varphi}[\mu] \subseteq [\varphi \wedge \mu]$ is entirely similar as for Theorem 3.3. To show that $[\varphi \wedge \mu] \subseteq \min_{\leq_\varphi}[\mu]$, suppose that there exists an interpretation $w_1 \in [\varphi \wedge \mu]$ such that $w_1 \notin \min_{\leq_\varphi}[\mu]$. This means that there exists $w_2 \in \min_{\leq_\varphi}[\mu]$ such that $w_2 <_\varphi w_1$. Using the previous result, we can conclude that $w_2 \in [\varphi]$. Since $w_1 \in [\varphi]$, this contradicts property r_6.

Theorems 3.1 and 3.3 add the final touches to a variant of revision that makes up the standard, received Katsuno-Mendelzon model. Stitching them together with Theorems 3.2 and 3.4 gives us the classical representation results found in the literature, the first of which is for total preorders.

Theorem 3.5 ([Katsuno and Mendelzon, 1992])

A revision operator $\circ$ satisfies postulates R_{1-6} if and only if there exists an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that satisfies properties r_{1-7} (i.e., that is total, syntax insensitive and r-faithful) and that represents the operator $\circ$.

The assignment representing an exhaustive operator $\circ$, we know from Theorems 3.1, is the exhaustive $\circ$-revealed assignment $\preccurlyeq^{\text{exh}}$, based on pairwise comparisons of interpretations. The second result is for partial preorders.

Theorem 3.6 ([Katsuno and Mendelzon, 1992])

A revision operator $\circ$ satisfies postulates R_{1-5} and R_{7-8} if and only if there exists an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that satisfies properties r_{1-2}, r_4 and r_{6-7} (i.e., that is partial, syntax insensitive and r-faithful) and that represents the operator $\circ$.

In this case, the assignment representing $\circ$ is the exclusive $\circ$-revealed assignment. Theo-

rems 3.5 and 3.6 gather together the insights of the Katsuno-Mendelzon model: revision according to the postulates given in this section, in either of the variants considered, amounts to choosing the best outcomes available, according to a ranking on outcomes that is biased by the agent's belief.

But where do such rankings come from?

Distance-based revision operators

In Section 2.3 we presented a general method for computing distances from a propositional formula φ to an interpretation w, using two ingredients: the first is a quasi-distance function d between interpretations, used to generate a tuple $(d(v,w))_{v\in[\varphi]}$ of distances between every model of φ and w, while the second ingredient is an aggregation function $\oplus$ used to aggregate the values in the tuple $(d(v,w))_{v\in[\varphi]}$ and generate the $(d,\oplus)$-induced distance $d^{\oplus}(\varphi,w)$ from φ to w. In this section we want to use this notions to rank interpretations relative to a formula φ. Thus, if d is a quasi-distance between interpretations, $\oplus$ is an aggregation function, φ is a propositional formula, the $(d,\oplus)$-*induced ranking* $\leq_{\varphi}^{d,\oplus}$ is defined, for any two interpretations w_1, w_2, as follows:

$$w_1 \leq_{\varphi}^{d,\oplus} w_2 \text{ if } d^{\oplus}(\varphi,w_1) \leq d^{\oplus}(\varphi,w_2).$$

Intuitively, w_1 is considered better than w_2 according to $\leq_{\varphi}^{d,\oplus}$ if w_1 is closer to φ than w_2, according to the measures uses. For this section, where we are focused on techniques used in the existing literature, we will assume that the aggregation function is min throughout. Thus, to rephrase things, if d is a quasi-distance between interpretations, the $(d,\min)$-induced ranking $\leq_{\varphi}^{d,\min}$ is obtained by taking, for any two interpretations w_1, w_2:

$$w_1 \leq_{\varphi}^{d,\min} w_2 \text{ if } \min(d(v,w_1))_{v\in[\varphi]} \leq \min(d(v,w_2))_{v\in[\varphi]}.$$

In other words, an agent whose prior belief is φ considers interpretation w_1 more plausible than w_2 if the shortest distance between w_1 and the models of φ is shorter than the shortest distance between w_2 and the models of φ, i.e., if w_1 is overall closer to the models of φ than w_2. If d is a quasi-distance, the $(d,\min)$-*induced assignment* $\preccurlyeq^{d,\min}$ is obtained by taking $\preccurlyeq^{d,\min}(\varphi) = \leq_{\varphi}^{d,\min}$, for any propositional formula φ. In the same vein, the $(d,\min)$-*induced revision operator* $\circ^{d,\min}$ is the operator induced by the assignment $\preccurlyeq^{d,\min}$. This allows us to generate total, syntax insensitive r-faithful assignments.

Proposition 3.2

If d is a quasi-distance between interpretations and φ is a propositional formula, the $(d,\min)$-induced ranking $\leq_{\varphi}^{d,\min}$ satisfies properties r_{1-5} and r_7, i.e., $\leq_{\varphi}^{d,\min}$ is total, syntax insensitive and r-faithful.

Proof

Since interpretations in $\leq_\varphi^{d,\,\min}$ are ranked based on the min-aggregated distance from φ, which in this case is a single number, it is straightforward to see that $\leq_\varphi^{d,\,\min}$ is a total preorder on interpretations, i.e., that $\leq_\varphi^{d,\,\min}$ satisfies properties r_{1-3}. Since the definition of $\leq_\varphi^{d,\,\min}$ depends only on the interpretations, $\leq_\varphi^{d,\,\min}$ also satisfies property r_4. Finally, it holds that $d^{\min}(\varphi, w) = \min(d(v, w))_{v\in[\varphi]} = 0$ if and only if $w \in [\varphi]$, which implies that models of φ are the $\leq_\varphi^{d,\,\min}$-minimal elements in $\leq_\varphi^{d,\,\min}$, i.e., $\leq_\varphi^{d,\,\min}$ satisfies properties r_5 and r_7, for any propositional formula φ.

Proposition 3.2 implies that the $(d,\,\min)$-induced assignment $\preccurlyeq^{d,\,\min}$ is total, r-faithful and syntax insensitive, which, by Theorem 3.5, implies that the $(d,\,\min)$-induced revision operator $\circ^{d,\,\min}$ satisfies postulates R_{1-6}.

Corollary 3.1

If d is a quasi-distance between interpretations, the $(d,\,\min)$-induced revision operator $\circ^{d,\,\min}$ satisfies postulates R_{1-6}.

The operators generated using Hamming distance d_{H} and drastic distance d_{D}, as presented in Section 2.3, are denoted $\circ^{\mathrm{H},\,\min}$ and $\circ^{\mathrm{D},\,\min}$, respectively. We will refer to $\circ^{\mathrm{H},:\min}$ as *Dalal's operator*, for historical reasons [Dalal, 1988], and to $\circ^{\mathrm{D},\,\min}$ as the *drastic operator*. Dalal's operator and the drastic operator are intuitive examples of revision operators that satisfy postulates R_{1-6}, but Corollary 3.1 shows us that they are just two instances of the much larger framework of $(d,\,\min)$-induced operators. What these operators all have in common is the choice procedure used to select the best outcomes and the fact that they are based on total preorders.

To get partial preorders, we use two quasi-distance functions d_1 and d_2, in addition to the min aggregation function. On the basis of this, the $((d_1, d_2),\,\min)$-*induced ranking* $\leq_\varphi^{(d_1,d_2),\,\min}$ on interpretations is defined, for any interpretations w_1 and w_2, by taking:

$$w_1 \leq_\varphi^{(d_1,d_2),\,\min} w_2 \text{ if } d_1^{\min}(\varphi, w_1) \leq d_1^{\min}(\varphi, w_2) \text{ and } d_2^{\min}(\varphi, w_1) \leq d_2^{\min}(\varphi, w_2).$$

Correspondingly, the $((d_1, d_2),\,\min)$-*induced assignment* $\preccurlyeq^{(d_1,d_2),\,\min}$ is obtained by taking $\preccurlyeq^{(d_1,d_2),\,\min}(\varphi) = \leq_\varphi^{(d_1,d_2),\,\min}$, for any propositional formula φ. In the same vein, the $((d_1, d_2),\,\min)$-*induced revision operator* $\circ^{(d_1,d_2),\,\min}$ is the operator induced by the assignment $\preccurlyeq^{(d_1,d_2),\,\min}$.

Proposition 3.3

If d_1 and d_2 are quasi-distances between interpretations and φ is a propositional formula, the $((d_1, d_2),\,\min)$-induced ranking $\leq_\varphi^{(d_1,d_2),\,\min}$ satisfies properties r_{1-2}, r_4,

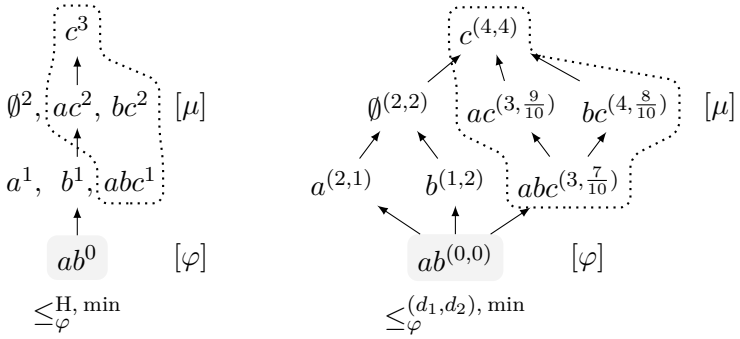


Figure 3.3: A total preorder $\leq_\varphi^{\mathrm{H},\,\min}$ and a partial preorder $\leq_\varphi^{(d_1,d_2),\,\min}$. The distances from φ to each interpretation are written as superscripts next to each interpretation. Models of φ are shaded in gray, models of μ are in the region bounded by the dotted border.

r_6 and r_7, i.e., $\leq_\varphi^{(d_1,d_2),\,\min}$ is partial, syntax insensitive and r-faithful.

Proof

It is straightforward to see that the (d_1, d_2), min-induced relation $\leq_\varphi^{(d_1,d_2),\,\min}$ is a partial preorder on $\mathcal{U}$ that is syntax insensitive and still puts models of φ on the bottom, i.e., that it satisfies properties r_{1-2}, r_4, r_6 and r_7, for any propositional formula φ.

Proposition 3.3 implies that the $((d_1, d_2)$, min)-induced assignment $\preccurlyeq^{(d_1,d_2),\,\min}$ is r-faithful, which, by Theorem 3.6, implies that the $((d_1, d_2)$, min)-induced revision operator $\circ^{(d_1,d_2),\,\min}$ satisfies postulates R_{1-5} and R_{7-8}.

Corollary 3.2

If d_1 and d_2 are quasi-distances between interpretations, the $((d_1, d_2)$, min)-induced revision operator $\circ^{(d_1,d_2),\,\min}$ satisfies postulates R_{1-5} and R_{7-8}.

An example will clarify the two main approaches.

Example 3.6: A monopoly on tool use no more

For the last time in this chapter, we look at the revision scenario from Example 3.1, for which $[\varphi] = \{ab\}$ and $[\mu] = \{c, ac, bc, abc\}$. The preorder $\leq_\varphi^{\mathrm{H},\,\min}$ generated using Hamming distance and the min aggregation function is depicted on the left in Figure

3.3. According to it we obtain that $[\varphi \circ^{\mathrm{H},\,\min} \mu] = \min_{\leq_\varphi^{\mathrm{H},\,\min}}[\mu] = \{abc\}$. Consider also two quasi-distances d_1 and d_2 between interpretations that generate the partial preorder $\leq_\varphi^{(d_1,d_2),\,\min}$ depicted on the right in Figure 3.3. According to it we obtain $[\varphi \circ^{(d_1,d_2),\,\min} \mu] = \min_{\leq_\varphi^{(d_1,d_2),\,\min}}[\mu] = \{abc\}$, which is the same result as for $\circ^{\mathrm{H},\,\min}$.

In both cases, the revision operators arrive at the same conclusion that has ultimately prevailed in the primatology community: the minimally disruptive response to Jane Goodall's findings is to hold on to the beliefs that humans use tools and that chimpanzees are a different species from humans, but to accept that chimpanzees can use tools.

We end this section by returning to a point that was made at its beginning. The point is that, at least insofar as postulates R_{1-8} are concerned, the type of entity represented by φ and μ should be conceived as fluid, hovering somewhere in the space of cognitive attitudes an agent can have towards a generic set of issues, but exclusive to neither of them in particular. This is seen more clearly through the choice lens, embodied by Theorems 3.5 and 3.6): postulates R_{1-8} axiomatize preference maximizing behavior, i.e., an operation that selects the best alternatives out of a set menu, biasing the judgment of what is best on the prior information available; this is behavior that is no more exclusive to beliefs than it is to actions or bundles of goods, and it can be expected to be part of a rational agent's arsenal in all of these cases. The general appeal of framing rational behavior in this way was understood early on in economics [Nash, 1950, Arrow, 1951, Chernoff, 1954, Radner and Marschak, 1954, Luce and Raiffa, 1957, Hansson, 1968, Sen, 1969, Sen, 1970, Herzberger, 1973], and is what lies, for Hans Rott, at the root of both theoretical and practical reason:

> The constraints [for rational or coherent choice] are shown to give rise to corresponding lists of conditions for [...] revision and inference operations. I take this to be strong evidence for the unity of theoretical and practical reason, with the principles for the former being special cases of principles for the latter. [Rott, 1992, p. 214]

As mentioned before, the moral we want to draw from here is not that postulates R_{1-8} are the last word in what constitutes rational behavior, but that they are parts of a larger framework that is worth exploring further.

3.2 Update

Revision, as we have seen in Section 3.1, works by choosing the best outcomes from the ones consistent with the new information, or, in what is the same thing, by discarding any outcomes from the new information that are not optimal. While this selection process makes sense in certain scenarios, there are cases in which it ends up being too aggressive.

Example 3.7: Keeping up with the humans, as an update task

Consider the scenario in Example 1.3. The variables that my automatic assistant keeps track of are whether the temperature is above 15° C (a), whether the Wi-Fi is on after 21:00 (b), and whether my friend is online after 21:00 (c). The instructions my assistant is programmed to implement are $\varphi = a \wedge \neg b$, whereas my observed pattern of behavior is represented by the formula $\mu = (b \leftrightarrow c)$. The assistant would like to modify its list of instructions to accommodate my behavior, i.e., to change φ in accordance with μ. In this case, it seems like the sensible answer is to move from φ to $\varphi' = a \wedge (b \leftrightarrow c)$, i.e., maintain temperature above 15° C and leave the Wi-Fi on after 21:00 at exactly those times when my friend is online. Note that revision is not the appropriate operation here: since φ and μ are consistent, a typical revision operator would return $\varphi \wedge \mu \equiv a \wedge \neg b \wedge \neg c$: according to the logic of revision my smarthome would infer that, since it is after 21:00 and the Wi-Fi is turned off, then my friend must be offline.

Example 3.7 illustrates the need for a belief change operator that retains more information from the new information μ than a revision operator would normally do, while still being biased by φ. The bias towards φ, therefore, should not be so strong as to render all but the absolute closest outcomes as unfeasible. Update operators were introduced to do justice to this intuition [Katsuno and Mendelzon, 1991], and we will see that they do so by modifying the way in which models of μ are chosen for the final result. In the rest of this section we will focus on the mechanics of update, using the same methodology as the one used for revision: postulates, preferences over outcomes, representation theorems and distances between interpretations.

Postulates

Like revision, update is a single-agent belief change operator. Formally, an *$\mathcal{L}$-update operator* $\diamond$ is a function $\diamond\colon \mathcal{L} \times \mathcal{L} \to \mathcal{L}$, taking as input two propositional formulas, denoted here by φ and μ, and standing in for the agent's prior information and the newly acquired information, respectively, and returning a propositional formula, denoted here by $\varphi \diamond \mu$, and standing for the agent's posterior information. As with revision, φ and μ are nominally intended to be beliefs, but in practice can be any of a number of cognitive attitudes an agent can have toward a set of items.

Recall that a complete formula $\dot{\varphi}$ is complete if $\dot{\varphi}$ has exactly one model. If $\diamond$ is an $\mathcal{L}$-update operator, the postulates $\diamond$ is expected to satisfy are, for any propositional formulas φ, φ_1, φ_2, complete formulas $\dot{\varphi}$, μ, μ_1 and μ_2, as follows:

(U_1) $\varphi \diamond \mu \models \mu$.

(U_2) If $\varphi \models \mu$, then $\varphi \diamond \mu \equiv \varphi$.

(U_3) If φ and μ are satisfiable, then $\varphi \diamond \mu$ is satisfiable.

(U$_4$) If $\varphi_1 \equiv \varphi_2$ and $\mu_1 \equiv \mu_2$, then $\varphi_1 \diamond \mu_1 \equiv \varphi_2 \diamond \mu_2$.

(U$_5$) $(\varphi \diamond \mu_1) \wedge \mu_2 \models \varphi \diamond (\mu_1 \wedge \mu_2)$.

(U$_6$) If $(\dot{\varphi} \diamond \mu_1) \wedge \mu_2$ is consistent, then $\dot{\varphi} \diamond (\mu_1 \wedge \mu_2) \models (\dot{\varphi} \diamond \mu_1) \wedge \mu_2$.

(U$_7$) If $\varphi \diamond \mu_1 \models \mu_2$ and $\varphi \diamond \mu_2 \models \mu_1$, then $\varphi \diamond \mu_1 \equiv \varphi \diamond \mu_2$.

(U$_8$) If $\mu \equiv \mu_1 \vee \mu_2$. then $(\dot{\varphi} \diamond \mu_1) \wedge (\dot{\varphi} \diamond \mu_2) \models \dot{\varphi} \diamond \mu$.

(U$_9$) $(\varphi_1 \vee \varphi_2) \diamond \mu \equiv (\varphi_1 \diamond \mu) \vee (\varphi_2 \diamond \mu)$.

As for revision, postulate U$_6$ implies postulates U$_7$ and U$_8$, and the plan with respect to their use is the same: postulates U$_{7-8}$ are meant to be alternatives to U$_6$. An update operator $\diamond$ is *exhaustive* if it satisfies postulates U$_{1-6}$ and U$_9$, and *exclusive* if it satisfies postulates U$_{1-5}$ and U$_{7-9}$.

The numbering of the postulates is slightly different from the usual ordering [Katsuno and Mendelzon, 1991], but the re-numbering is meant to highlight the close connection to revision. Indeed, note that postulates U$_1$ and U$_{3-8}$ are esentially similar to the revision postulates R$_1$ and R$_{3-8}$ (see Section 3.1), with the only point of departure being that postulates U$_6$ and U$_8$ are meant to apply only to complete formulas $\dot{\varphi}$. What is more, if φ is a complete formula, then postulates U$_{1-8}$ are entirely equivalent to revision postulates R$_{1-8}$: this is true even for postulate U$_2$, since if φ is complete then $\varphi \wedge \mu$ and φ become equivalent and the statement that $\varphi \models \mu$ is equivalent to the statement that $\varphi \wedge \mu$ is consistent. This is an observation that has been made before [Peppas et al., 1996] but is worth stressing, since it provides insight into the working of an update operator: on complete propositional formulas, update according to postulates U$_{1-8}$ is just revision according to postulates R$_{1-8}$.

If φ is not complete, then postulates U$_2$ and U$_9$ kick in, and can be seen as new additions to the toolbox of familiar postulates. In this case postulate U$_2$ is a weaker version of the revision postulate R$_2$, and regulates the way in which the prior information φ biases the update result. Postulate U$_9$ specifies the way in which the update result can be decomposed in results for more specific parts of the prior information φ, i.e., formulas φ_1 and φ_2 such that $\varphi_1 \vee \varphi_2 \equiv \varphi$. Ultimately, repeated application of postulate U$_9$ makes the result for $\varphi \diamond \mu$ entirely dependent on the update result for the complete formulas that imply φ. More precisely, if ε_v is an $\mathcal{L}$-proxy for the interpretation v, i.e., a propositional formula such that $[\varepsilon_v] = \{v\}$, then postulate U$_9$ is equivalent to the following postulate, applying for any propositional formulas φ and μ:

(U$_{10}$) $\varphi \diamond \mu \equiv \bigvee_{v \in [\varphi]} (\varepsilon_v \diamond \mu)$.

Postulate U$_{10}$ shows that $\varphi \diamond \mu$ can be decomposed in the results for $\varepsilon_v \diamond \mu$, for every $v \in [\varphi]$. We will make extensive use of postulate U$_9$ and, even more so, of its equivalent reformulation U$_{10}$, in what is to follow.

Example 3.8: Update is not revision

For the setting in Example 3.7, where $\varphi = a \wedge \neg b$ and $\mu = b \leftrightarrow c$, we have that $\varphi \wedge \mu$ is consistent, but $\varphi \not\models \mu$. Thus, whereas a revision operator satisfying postulate R_2 would require the result to be $\varphi \wedge \mu$, the update postulate U_2 places no constraints in this case.

Since $[\varphi] = \{a, ac\}$, postulate U_9 (or U_{10}) requires that $\varphi \diamond \mu \equiv (\varepsilon_a \diamond \mu) \vee (\varepsilon_{ac} \diamond \mu)$.

As with the revision postulate R_8 in Section 3.1, postulate U_8 has also been slightly re-phrased: normally there would be no reference to μ, with $\mu_1 \vee \mu_2$ written instead. But here, as well, the difference from the usual statement is merely stylistic. The role of this re-phrasing is only to make life easier in Chapter 6.

Preferences over outcomes

Postulate U_9, and even more so postulate U_{10}, show that what gets chosen in $\varphi \diamond \mu$ is determined by what gets chosen in $\varepsilon_v \diamond \mu$, for every $v \in [\mu]$. Furthermore, update for complete formulas is just revision. This provides a useful hint for how to model update as a choice procedure: we will use an $\mathcal{L}_{\text{comp}}$*-assignment* $\preccurlyeq$ *on interpretations*, which is a function $\preccurlyeq\colon \mathcal{L}_{\text{comp}} \to 2^{\mathcal{U} \times \mathcal{U}}$, taking as input a complete formula $\dot{\varphi}$ and returning a binary relation on interpretations, interpreted, as for revision, as a plausibility ranking. An $\mathcal{L}_{\text{comp}}$ assignment corresponds to what has been called in the literature as a *pointwise assignment* [Katsuno and Mendelzon, 1991].

As expected, we are keen on $\preccurlyeq$ satisfying some desirable properties, and the properties we are interested in are as follows, for any complete propositional formulas $\dot{\varphi}$, $\dot{\varphi}_1$, $\dot{\varphi}_2$ and interpretations w, v, w_1 and w_2:

(u_1) $w \leq_{\dot{\varphi}} w$.

(u_2) If $w_1 \leq_{\dot{\varphi}} w_2$ and $w_2 \leq_{\dot{\varphi}} w_3$, then $w_1 \leq_{\dot{\varphi}} w_3$.

(u_3) $w_1 \leq_{\dot{\varphi}} w_2$ or $w_2 \leq_{\dot{\varphi}} w_1$.

(u_4) If $\dot{\varphi}_1 \equiv \dot{\varphi}_2$, then it holds that if $w_1 \leq_{\dot{\varphi}_1} w_2$, then $w_1 \leq_{\dot{\varphi}_2} w_2$.

(u_5) If $[\dot{\varphi}] = \{v\}$ and $w \neq v$, then $v <_{\dot{\varphi}} w$.

Properties u_{1-4} are the same as properties r_{1-4}, except that they are particularized to complete formulas. They say, in effect, that $\leq_{\dot{\varphi}}$ is a preorder (properties u_{1-2}), additionally total (property u_3) and syntax insensitive (property u_4), for any complete propositional formula $\dot{\varphi}$. Property u_5 conveys the same message as properties r_{5-7}: models of $\dot{\varphi}$ are the unique minimal elements in $\leq_{\dot{\varphi}}$, but, as $\dot{\varphi}$ has only one model, the

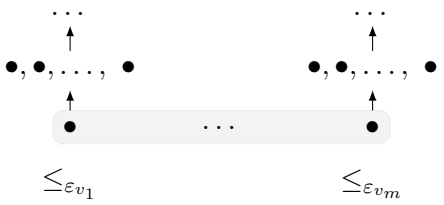


Figure 3.4: A schematic depiction of total preorders $\leq_{\varepsilon_{v_i}}$, for $[\varphi] = \{v_1, \ldots, v_m\}$, in a total u-faithful assignment. Bullets stand, as before, for interpretations. Each model of φ (placed in the shaded gray region) generates its own total preorder on interpretations.

distinctions inherent in properties r_{5-7} are not needed. We are left, in this case, with the simpler property u_5.

An $\mathcal{L}_{\text{comp}}$-assignment $\preccurlyeq$ on interpretations is *partial* if it satisfies properties u_{1-2}, *total* if it satisfies properties u_{1-3}, syntax insensitive if it satisfies property u_4 and *u-faithful* if it satisfies property u_5. A schematic illustration of preorders in a total u-faithful assignment is given in Figure 3.4.

Update as choice over outcomes

Seeing update as a choice procedure over outcomes involves putting together the two perspectives introduced in the previous paragraphs: the logical postulates, on the one side, and the plausibility rankings on interpretations, on the other. As with revision, plausibility rankings can be used to guide update, as well as be inferred from update behavior.

Thus, keeping in mind that ε_v is a propositional formula such that $[\varepsilon_v] = \{v\}$, then, given an $\mathcal{L}_{\text{comp}}$-assignment $\preccurlyeq$ on interpretations, the *$\preccurlyeq$-induced update operator* $\diamond^{\preccurlyeq}$ is defined, for any propositional formulas φ and μ, by taking:

$$[\varphi \diamond^{\preccurlyeq} \mu] \stackrel{\text{def}}{=} \bigcup_{v \in [\varphi]} \min_{\leq_{\varepsilon_v}} [\mu].$$

Conversely, given an $\mathcal{L}$-update operator $\diamond$, and a complete propositional formula $\dot{\varphi}$, the *exhaustive $\diamond$-revealed plausibility relation* $\leq^{\text{exh}}_{\dot{\varphi}}$ and the *exclusive $\diamond$-revealed plausibility relation* $\leq^{\text{exc}}_{\dot{\varphi}}$ are defined, for any interpretations w_1 and w_2, respectively, as:

$$w_1 \leq^{\text{exh}}_{\dot{\varphi}} w_2 \text{ if } w_1 \in [\dot{\varphi} \circ \varepsilon_{1,2}],$$
$$w_1 \leq^{\text{exc}}_{\dot{\varphi}} w_2 \text{ if } w_1 \in [\dot{\varphi} \circ \varepsilon_{1,2}] \text{ and } w_2 \notin [\dot{\varphi} \circ \varepsilon_{1,2}].$$

The *exhaustive revealed $\mathcal{L}_{\text{comp}}$-assignment* $\preccurlyeq^{\text{exh}}$ and *exclusive revealed $\mathcal{L}_{\text{comp}}$-assignment* $\preccurlyeq^{\text{exc}}$ are obtained by taking $\preccurlyeq^{\text{exh}}(\dot{\varphi}) = \leq^{\text{exh}}_{\dot{\varphi}}$ and $\preccurlyeq^{\text{exc}}(\dot{\varphi}) = \leq^{\text{exc}}_{\dot{\varphi}}$, for any complete propositional formula $\dot{\varphi}$. The guiding intuition here is the same as for the exhaustive and exclusive revealed assignments in revision.

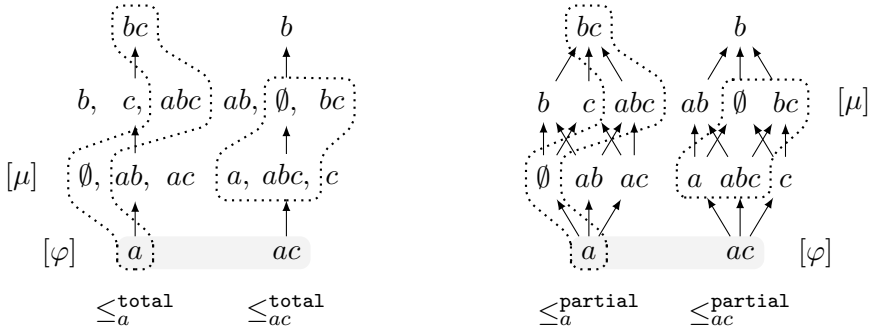


Figure 3.5: Total and partial preorders $\leq_{\varepsilon_v}^{\texttt{total}}$ and $\leq_{\varepsilon_v}^{\texttt{partial}}$, for $[\varphi] = \{a, ac\}$ and $v \in [\varphi]$, assigned by a total u-faithful assignment $\preccurlyeq^{\texttt{total}}$ and a partial u-faithful assignment $\preccurlyeq^{\texttt{partial}}$, respectively. Models of φ are in the shaded gray regions. The new information is μ, with $[\mu] = \{\emptyset, a, bc, abc\}$. The result of updating φ by μ using the assignment $\preccurlyeq^{\texttt{total}}$ amounts to taking the best models of μ from the preorders associated to each model of φ, i.e., from $\leq_a^{\texttt{total}}$ and from $\leq_{ac}^{\texttt{total}}$. The same strategy applies to $\preccurlyeq^{\texttt{partial}}$.

Example 3.9: Keeping up with the humans, using assignments

For the setting in Example 3.7, with $[\varphi] = \{a, ac\}$ and $[\mu] = \{\emptyset, a, bc, abc\}$, consider two assignments: a total r-faithful assignment $\preccurlyeq^{\texttt{total}}$ and a partial r-faithful assignment $\preccurlyeq^{\texttt{partial}}$, generating the preorders $\leq_{\varepsilon_v}^{\texttt{total}}$ and $\leq_{\varepsilon_v}^{\texttt{partial}}$ in Figure 3.5, for $v \in [\varphi]$. These assignments generate the update operators $\diamond^{\texttt{total}}$ and $\diamond^{\texttt{partial}}$, respectively, according to which:

$$\begin{aligned}[\varphi \diamond^{\texttt{total}} \mu] &= \min_{\leq_a^{\texttt{total}}}[\mu] \cup \min_{\leq_{ac}^{\texttt{total}}}[\mu] \\ &= \{a, abc\} \\ &= \min_{\leq_a^{\texttt{partial}}}[\mu] \cup \min_{\leq_{ac}^{\texttt{partial}}}[\mu] \\ &= [\varphi \diamond^{\texttt{partial}} \mu].\end{aligned}$$

Conversely, to find out the agent's ranking of, say, outcomes b and c, if its initial belief were the complete formula ε_a, with $[\varepsilon_a] = \{a\}$, we would look at the result of $\varepsilon_a \diamond \varepsilon_{b,c}$. Supposing that the result is $[\varepsilon_a \diamond \varepsilon_{b,c}] = \{b, c\}$, then according to the exhaustive $\diamond$-revealed $\mathcal{L}_{\text{comp}}$-assignment, we would conclude that $b \approx_{\varepsilon_a}^{\text{exh}} c$, whereas according to the exclusive $\diamond$-revealed $\mathcal{L}_{\text{comp}}$-assignment we would conclude that neither $b \leq_{\varepsilon_a}^{\text{exc}} c$ nor $c \leq_{\varepsilon_a}^{\text{exc}} b$. These results are in accordance with $\leq_{\varepsilon_a}^{\texttt{total}}$ and $\leq_{\varepsilon_a}^{\texttt{partial}}$ as in Figure 3.5.

If $\diamond$ is an $\mathcal{L}$-update operator and $\preccurlyeq$ is an $\mathcal{L}_{\text{comp}}$-assignment on interpretations, then $\preccurlyeq$ *represents* $\diamond$ (and $\diamond$ *is represented by* $\preccurlyeq$) if, for any propositional formulas φ and μ, it holds that $[\varphi \diamond \mu] = \bigcup_{v \in [\varphi]} \min_{\leq_{\varepsilon_v}}[\mu]$.

As with revision, we obtain two representation results for update operators satisfying either postulates U_{1-6} and U_9, or postulates U_{1-5}, U_{7-9}, one for total preorders and one for partial preorders. The first result is in terms of total preorders

Theorem 3.7 ([Katsuno and Mendelzon, 1991])

An update operator $\diamond$ satisfies postulates U_{1-6} and U_9 (i.e., is exhaustive) if and only if there exists an $\mathcal{L}_{\text{comp}}$-assignment $\preccurlyeq$ on interpretations that satisfies properties u_{1-5} (i.e., is total, syntax insensitive and u-faithful) and that represents the operator $\diamond$.

Proof

The $\mathcal{L}_{\text{comp}}$-assignment representing $\diamond$ is exactly the $\diamond$-revealed exhaustive assignment $\preccurlyeq^{\text{exh}}$, and the proof that it satisfies properties u_{1-5} and that $[\dot{\varphi} \diamond \mu] = \min_{\leq^{\text{exh}}_{\dot{\varphi}}}[\mu]$, for any formula μ and complete formula $\dot{\varphi}$, is essentially similar to the proof for the exhaustive revealed assignment of a revision operator (see Theorems 3.1 and 3.3). For the last step, i.e., showing that $[\varphi \diamond \mu] = \bigcup_{v \in [\varphi]} \min_{\leq^{\text{exh}}_{\varepsilon_v}}[\mu]$, postulate U_9 (or, more precisely, postulate U_{10}) is used.

The accompanying result trades postulate U_6 for U_{7-8} to obtain partial preorders instead of total preorders.

Theorem 3.8 ([Katsuno and Mendelzon, 1991])

An update operator $\diamond$ satisfies postulates U_{1-5}, U_{7-8} and U_9 (i.e., is exclusive) if and only if there exists an $\mathcal{L}_{\text{comp}}$-assignment $\preccurlyeq$ on interpretations that satisfies properties u_{1-2} and u_{4-5} (i.e., is partial, syntax insensitive and u-faithful) and that represents the operator $\diamond$.

Proof

The $\mathcal{L}_{\text{comp}}$-assignment representing $\diamond$ is exactly the $\diamond$-revealed exclusive assignment $\preccurlyeq^{\text{exc}}$, and the proof that it satisfies properties u_{1-2} and u_{4-5} and that $[\dot{\varphi} \diamond \mu] = \min_{\leq^{\text{exc}}_{\dot{\varphi}}}[\mu]$, for any formula μ and complete formula $\dot{\varphi}$, is essentially similar to the proof for the exclusive revealed assignment of a revision operator (see Theorems 3.2 and 3.4). For the last step, i.e., showing that $[\varphi \diamond \mu] = \bigcup_{v \in [\varphi]} \min_{\leq^{\text{exc}}_{\varepsilon_v}}[\mu]$, postulate U_9 (or, more precisely, postulate U_{10}) is used.

As with revision, we can refine the analysis by separating postulate U_2 from the rest of the postulates: what we obtain, then, are update operators represented by either total or partial $\mathcal{L}_{\text{comp}}$-assignments that are syntax insensitive but do not, however, satisfy property u_5 (i.e., are not u-faithful). In other words, postulate U_2 regulates, as for

revision, the position of the model of a complete formula $\dot{\varphi}$ in $\leq_{\dot{\varphi}}$ and its absence means that this model can be placed anywhere in $\leq_{\dot{\varphi}}$.

Total and partial preorders

Theorems 3.7 and 3.8 tell us that to obtain update operators that satisfy postulates U_{1-9} we must look for ways of generating rankings on interpretations. These rankings are supposed to depend on a single interpretation, so the tried and tested method of using a quasi-distance d together with an aggregation function $\oplus$ works here smoothly (see Section 2.3). As for revision, we will take $\oplus$ in this section to be the min aggregation function. Thus, if d is a quasi-distance between interpretations and v, w_1 and w_2 are interpretations, the $(d,\ \min)$-induced ranking $\leq_{\varepsilon_v}^{d,\ \min}$ is obtained by taking:

$$w_1 \leq_{\varepsilon_v}^{d,\ \min} w_2 \text{ if } d(v, w_1) \leq d(v, w_2).$$

If d is a quasi-distance, the $(d,\ \min)$-*induced assignment* $\preccurlyeq^{d,\ \min}$ is obtained, similarly as for revision, by taking $\preccurlyeq^{d,\ \min}(\dot{\varphi}) = \leq_{\dot{\varphi}}^{d,\ \min}$, for any complete propositional formula $\dot{\varphi}$. The $(d,\ \min)$-*induced* $\mathcal{L}$-*update operator* $\diamond^{d,\ \min}$ is the operator induced by the $\mathcal{L}_{\mathrm{comp}}$-assignment $\preccurlyeq^{d,\ \min}$. The assignments that are generated in this way turn out to be total, syntax insensitive and u-faithful.

Proposition 3.4

If d is a quasi-distance between interpretations and $\dot{\varphi}$ is a complete propositional formula, the $(d,\ \min)$-induced ranking $\leq_{\dot{\varphi}}^{d,\ \min}$ satisfies properties r_{1-5}, i.e., $\leq_{\dot{\varphi}}^{d,\ \min}$ is total, syntax insensitive and u-faithful.

Proof

Entirely similar to the proof for Proposition 3.2.

Proposition 3.4 implies that the $(d,\ \min)$-induced assignment $\preccurlyeq^{d,\ \min}$ is total, u-faithful and syntax insensitive, which, by Theorem 3.7, implies that the $(d,\ \min)$-induced update operator $\diamond^{d,\ \min}$ satisfies postulates U_{1-9}.

Corollary 3.3

If d is a quasi-distance between interpretations, the $(d,\ \min)$-induced update operator $\diamond^{d,\ \min}$ satisfies postulates U_{1-6} and U_9.

The operators generated using Hamming distance d_{H} and drastic distance d_{D} are denoted $\diamond^{\mathrm{H},\ \min}$ and $\diamond^{\mathrm{D},\ \min}$, respectively. We will refer to $\diamond^{\mathrm{H},\ \min}$ as *Forbus's operator* [Forbus, 1989], and to $\diamond^{\mathrm{D},\ \min}$ as the *drastic update operator*.

Since any update operator that satisfies postulates U_{1-6} also satisfies postulates U_{7-8}, Forbus's operator and the drastic update operator work as examples for both exhaustive and exclusive operators. To obtain purely exclusive operators we could use the same technique as in Section 3.1, i.e., employ two quasi-distances d_1 and d_2. However, we will present a different type of operator here. If w_1 and w_2 are two interpretations, the *symmetric difference* $w_1 \triangle w_2$ *between* w_1 *and* w_2 is defined as:

$$w_1 \triangle w_2 = (w_1 \setminus w_2) \cup (w_2 \setminus w_1).$$

We have already used the cardinality of the symmetric difference between w_1 and w_2 to define the Hamming distance between w_1 and w_2, but here we will focus on the actual contents of the symmetric difference. Thus, if v, w_1 and w_2 are interpretations, the `symdiff`-induced ranking $\leq_{\varepsilon_v}^{\texttt{symdiff}}$ is obtained by taking:

$$w_1 \leq_{\varepsilon_v}^{\texttt{symdiff}} w_2 \text{ if } (v \triangle w_1) \subseteq (v \triangle w_2).$$

If d is a quasi-distance, the `symdiff`-*induced* $\mathcal{L}_{\text{comp}}$-*assignment* $\preccurlyeq^{\texttt{symdiff}}$ is obtained, by taking $\preccurlyeq^{\texttt{symdiff}}(\dot{\varphi}) = \leq_{\dot{\varphi}}^{\texttt{symdiff}}$, for any complete propositional formula $\dot{\varphi}$. The `symdiff`-*induced* $\mathcal{L}$-*update operator* $\diamond^{\texttt{symdiff}}$, is the operator induced by the $\mathcal{L}_{\text{comp}}$-assignment $\preccurlyeq^{\texttt{symdiff}}$. The `symdiff`-induced $\mathcal{L}$-update operator $\diamond^{\texttt{symdiff}}$ is also called Winslett's operator [Winslett, 1990].

The $\mathcal{L}_{\text{comp}}$-assignment $\preccurlyeq^{\texttt{symdiff}}$ turns out to be partial, syntax insensitive and u-faithful.

Proposition 3.5

If $\dot{\varphi}$ is a complete propositional formula, the `symdiff`-induced ranking $\leq_{\dot{\varphi}}^{\texttt{symdiff}}$ satisfies properties u_{1-2} and u_{4-5}, i.e., $\leq_{\dot{\varphi}}^{\texttt{symdiff}}$ is a partial preorder on interpretations that is syntax insensitive and makes the model of $\dot{\varphi}$ the $\leq_{\dot{\varphi}}^{\texttt{symdiff}}$-minimal element.

Proof

It is straightforward to see that $\leq_{\dot{\varphi}}^{\texttt{symdiff}}$ satisfies properties u_{1-2} and u_4. Notice, now, that if w is an interpretation such that $v \neq w$, then $v \triangle w \neq \emptyset$, whereas $v \triangle v = \emptyset$. Thus, for any $w \neq v$, it holds that $v <_{\varepsilon_v}^{\texttt{symdiff}} w$, which shows that $\leq_{\varepsilon_v}^{\texttt{symdiff}}$ satisfies property u_5.

Proposition 3.5 implies that the `symdiff`-induced assignment $\preccurlyeq^{\texttt{symdiff}}$ is partial, u-faithful and syntax insensitive, which, by Theorem 3.8, implies that the `symdiff`-induced update operator $\diamond^{\texttt{symdiff}}$ satisfies postulates U_{1-5} and U_{7-9}.

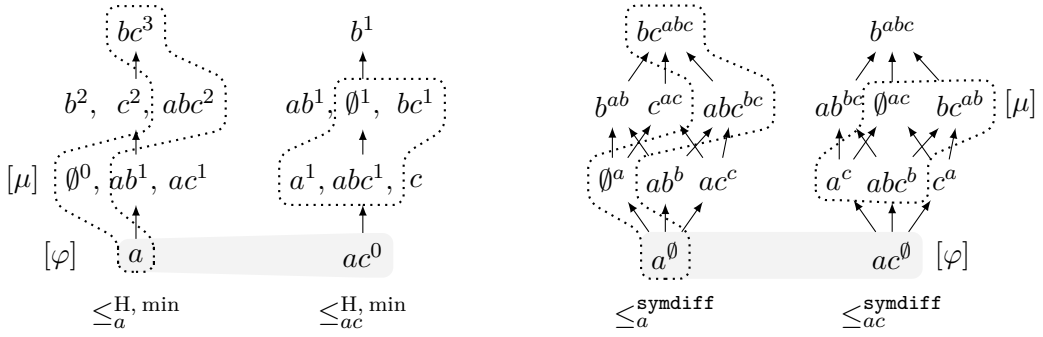


Figure 3.6: Total and partial preorders $\leq_{\varepsilon_v}^{\mathrm{H,\,min}}$ and $\leq_{\varepsilon_v}^{\mathtt{symdiff}}$, for $[\varphi] = \{a, ac\}$ and $v \in [\varphi]$, assigned by the assignments $\preccurlyeq^{\mathrm{H,\,min}}$ and $\preccurlyeq^{\mathtt{symdiff}}$, respectively. For $\preccurlyeq^{\mathrm{H,\,min}}$ distances are written as superscripts, whereas for $\preccurlyeq^{\mathtt{symdiff}}$ the symmetric differences are written as superscripts. Models of φ are in the shaded gray regions. The new information is μ, with $[\mu] = \{\emptyset, a, bc, abc\}$.

Corollary 3.4

The `symdiff`-induced update operator $\diamond^{\mathtt{symdiff}}$ satisfies postulates U_{1-6} and U_9.

Example 3.10: Keeping up with the humans, using distances

For the setting in Example 3.7, with $[\varphi] = \{a, ac\}$ and $[\mu] = \{\emptyset, a, bc, abc\}$, the (H, min)-induced assignment $\preccurlyeq^{\mathrm{H,\,min}}$ and the `symdiff`-induced assignment $\preccurlyeq^{\mathtt{symdiff}}$ generate the preorders in Figure 3.6. Note that $\leq_{\varepsilon_v}^{\mathrm{H,\,min}}$ and $\leq_{\varepsilon_v}^{\mathtt{symdiff}}$, for $v \in [\varphi]$, are the same as in Figure 3.5, and hence we obtain the same results for $\varphi \diamond^{\mathrm{H,\,min}} \mu$ and for $\varphi \diamond^{\mathtt{symdiff}} \mu$ as for $\varphi \diamond^{\mathtt{total}} \mu$ and $\varphi \diamond^{\mathtt{partial}} \mu$ in Example 3.9, respectively.

3.3 Enforcement

Both revision, as described in Section 3.1 and update, as described in Section 3.2, are based on the idea that new information is entirely trustworthy: even more trustworthy than prior information, to the point where if the two come into conflict, the new information has priority over any piece of prior information. Of course, this type of assumption is not always warranted: the agent might assess the source of the new information as equally reliable as its own elief formation process, such that new information may be considered plausible enough to be adopted as part of the agent's belief, but not necessarily more plausible than prior information. The challenge, then, would be to find place for new information alongside the old beliefs, but without necessarily dislodging them.

Example 3.11: The art of diagnosis as an enforcement problem

The scenario in Example 1.4 can be modeled by using propositional variables to represent the possible outcomes: allergic reaction (a), bronchitis (b) and the new strand of coronavirus (c). The doctor's initial belief φ is that the patient has an allergic reaction or bronchitis, i.e., $\varphi = a \vee b$. The patient's input μ is that it could also be the coronavirus, i.e., $\mu = c$. The doctor is willing to take this possibility into account, but does not think it more likely than the other two, and changes its belief to $\varphi \vee \mu = a \vee b \vee c$.

At the same time, if the patient had said: "I've been to another doctor and they told me it's neither an allergic reaction nor bronchitis", i.e., $\mu' = \neg a \wedge \neg b$, then the doctor might not be inclined to conclude $\varphi \vee \mu'$, which, in this case, is a tautology and it amounts to saying it could be anything. In such a case, the doctor might want to take that information into account, while not entirely discarding its own initial assessment.

Note that neither revision nor update is warranted in this case, since they both prescribe accepting μ.

Example 3.11 shows the need for an operation that can be thought of as a softer type of belief change than either revision or update, attempting to add as much information as possible to the store of existing beliefs and stopping short only of obtaining a tautology, and the enforcement operation we look at in this section captures exactly this type of change.

The idea that the new information should not be accepted without any reservation is not new to belief change, with much work in non-prioritized revision dedicated to formulating acceptable models of belief change in which this assumption is relaxed [Hansson, 1999a, Hansson et al., 2001]. However, none of the existing work on non-prioritized revision precisely captures the dynamics we have in mind here, so that enforcement as we put it forward is distinct from other existing types of belief change. The idea of enforcement can be traced back to previous publications on the dynamics of desire [Dubois et al., 2017], but the current section is based on work on *propositional enforcement* [Haret et al., 2018c], originally developed as an attempt to model enforcement in abstract argumentation [Baumann, 2012, Wallner et al., 2017], with the latter application providing inspiration for the name. Here we put propositional enforcement forward as a change operation in its own right, meant to stand alongside revision, update and the other members of the belief change family.

Postulates

An $\mathcal{L}$-*enforcement operator* $\triangleright$ is a function $\triangleright\colon \mathcal{L} \times \mathcal{L} \to \mathcal{L}$ that, like revision and update, takes propositional formulas φ and μ as input and produces a propositional formula $\varphi \triangleright \mu$ as output. Enforcement is a single-agent belief change operation and, following the

convention established for revision and update, we call φ, μ and $\varphi \triangleright \mu$ the prior, new and posterior information, respectively.

The postulates specific to enforcement apply for any propositional formulas φ, φ_1, φ_2 and μ, μ_1 and μ_2:

(E_1) $\mu \models \varphi \triangleright \mu$.

(E_2) If $\varphi \vee \mu$ is refutable, then $\varphi \triangleright \mu \equiv \varphi \vee \mu$.

(E_3) If μ is refutable, then $\varphi \triangleright \mu$ is refutable.

(E_4) If $\varphi_1 \equiv \varphi_2$ and $\mu_1 \equiv \mu_2$, then $\varphi_1 \triangleright \mu_1 \equiv \varphi_2 \triangleright \mu_2$.

(E_5) $\varphi \triangleright (\mu_1 \vee \mu_2) \models (\varphi \triangleright \mu_1) \vee \mu_2$.

(E_6) If $(\varphi \triangleright \mu_1) \vee \mu_2$ is refutable, then $(\varphi \triangleright \mu_1) \vee \mu_2 \models \varphi \triangleright (\mu_1 \vee \mu_2)$.

Postulate E_1 says that the newly acquired information μ should imply the enforcement result $\varphi \triangleright \mu$, which, in semantic terms, means that the outcomes consistent with μ are among the models of $\varphi \triangleright \mu$. If $\varphi \triangleright \mu$ is taken to encode the agents' epistemic state (i.e., the outcomes that are, in some sense, given priority), then postulate E_1 ensures that the models of μ are added to this set, i.e., that they are incorporated into the new epistemic state but not necessarily given priority over other interpretations. In accommodating μ with respect to φ, the simplest solution is to return, if possible, the disjunction $\varphi \vee \mu$, and this is exactly what postulate E_2 says. The success condition specifices that this should be done only if $\varphi \vee \mu$ is not a tautology, the reason being that a tautology carries no useful information and is best avoided, with postulate E_3 pushing this point. What to do, though, if $\varphi \vee \mu$ is a tautology? In this case postulates E_{1-3} provide no definite answer, only general guidelines: return a refutable formula implied by μ. What formula? The final answer is, again, a matter of choice and, as we have seen, choice must behave consistently across varying contexts, hence postulates E_{5-6}. Weaker versions of postulate E_6 can be considered, along the lines of revision postulates R_{7-8}, but to keep things clear and simple we will refrain from doing so here. Finally, postulate E_4 provides the usual insensitivity to the syntax of φ and μ.

Example 3.12: Possible results to an enforcement task

For φ, μ and μ' as in Example 3.11, we have that $\varphi \vee \mu = a \vee b \vee c$, which is a refutable formula. Thus, if $\triangleright$ is an enforcement operator satisfying postulates E_{1-6}, the result is $\varphi \triangleright \mu \equiv a \vee b \vee c$. On the other hand, $\varphi \vee \mu' \equiv \top$, and is not a valid answer. Postulates E_{1-4} require, in this case, that $[\neg a \wedge \neg b] \subseteq [\varphi \triangleright \mu] \subset \mathcal{U}$.

Contemplation of postulates E_{1-6} reveals that they can be obtained from the revision postulates R_{1-6} by replacing conjunction with disjunction and reversing the terms of

the implications. This similarity is not accidental, as enforcement turns out to be a sort of mirror image of revision. Concretely, given an $\mathcal{L}$-revision operator $\circ$, the $\circ$-*induced* $\mathcal{L}$-*enforcement operator* $\triangleright^{\circ}$ is defined, for any propositional formulas φ and μ, as:

$$\varphi \triangleright^{\circ} \mu \stackrel{\text{def}}{=} \neg(\neg\varphi \circ \neg\mu). \tag{3.1}$$

Interestingly, if $\circ$ is an $\mathcal{L}$-revision operator satisfying postulates R_{1-6} then the $\circ$-induced $\mathcal{L}$-enforcement operator $\triangleright^{\circ}$ turns out to satisfy postulates E_{1-6}.

Proposition 3.6 ([Haret et al., 2018c])

If $\circ$ is a revision operator satisfying postulates R_{1-6}, then the $\circ$-induced $\mathcal{L}$-enforcement operator $\triangleright^{\circ}$ satisfies postulates R_{1-6}.

Proof

Consider a revision operator $\circ$ satisfying postulates R_{1-6}. We will show that the $\circ$-induced $\mathcal{L}$-enforcement operator $\triangleright^{\circ}$ satisfies postulates E_{1-6}

For postulate E_1, we use postulate R_1 to get that $\neg(\neg\varphi \circ \neg\mu) \models \neg\mu$, which implies that $\mu \models \neg\varphi \circ \neg\mu$. For postulate E_2, notice that if $\mu \models \varphi$ then $\neg\varphi \models \neg\mu$. Since φ is assumed to be refutable, then $\neg\varphi$ is consistent, so $\neg\varphi \wedge \neg\mu$ is also consistent. Then, by postulate R_2, we have that $\neg\varphi \circ \neg\mu = \neg\varphi \wedge \neg\mu = \neg\varphi$, and hence $\neg(\neg\varphi \circ \neg\mu) = \varphi$. For postulate E_3, notice that if μ is refutable, then $\neg\mu$ is consistent and, by postulate R_3, it follows that $\neg\varphi \circ \neg\mu$ is consistent, hence $\neg(\neg\varphi \circ \neg\mu)$ is refutable. Postulate R_4 is immediate. For postulate E_5, apply R_5 to get that $(\neg\varphi \circ \neg\mu_1) \wedge \neg\mu_2 \models \neg\varphi \circ (\neg\mu_1 \wedge \neg\mu_2)$, which implies that $\neg(\neg\varphi \circ \neg(\mu_1 \vee \mu_2)) \models \neg(\neg\varphi \circ \neg\mu_1) \vee \mu_2$. For postulate E_6, notice that if $\neg(\neg\varphi \circ \neg\mu_1) \vee \mu_2$ is refutable, then $(\neg\varphi \circ \neg\mu_1) \wedge \neg\mu_2$ is consistent. We can thus apply postulate R_6 and get that $\neg\varphi \circ (\neg\mu_1 \wedge \neg\mu_2) \models (\neg\varphi \circ \neg\mu_1) \wedge \neg\mu_2$, which implies that $\neg(\neg\varphi \circ \neg\mu_1) \vee \mu_2 \models \neg(\neg\varphi \circ \neg(\mu_1 \vee \mu_2))$.

By entirely similar reasoning, an enforcement operator $\triangleright$ satisfying postulates E_{1-6} also induces a revision operator $\circ^{\triangleright}$ satisfying postulates R_{1-6}, called the $\triangleright$-*induced* $\mathcal{L}$-*revision operator* $\circ^{\triangleright}$, using the same maneuver:

$$\varphi \circ^{\triangleright} \mu \stackrel{\text{def}}{=} \neg(\neg\varphi \triangleright \neg\mu). \tag{3.2}$$

Equations 3.1 and 3.2 show that, at least at the syntactic level, we can switch between enforcement and revision whenever needed, while staying within the limits of postulates E_{1-6} and R_{1-6}. How do things look at the semantic level?

Preferences over outcomes

Sections 3.1 and 3.2 have primed us to expect that enforcement can be characterized as some sort of choice function over interpretations, with postulates E_{1-6} exploiting

a preference, or plausibility, relation on the interpretations themselves. The duality between enforcement and revision highlighted in the preceding paragraphs serves only to re-enforce this expectation. The first question, then, is what kind of properties should this putative plausibility relation satisfy.

We will use, as for revision, an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations, expected to satisfy the following properties, for any propositional formulas φ, φ_1 and φ_2 and interpretations w, w_1, w_2 and w_3:

(e_1) $w \leq_\varphi w$.

(e_2) If $w_1 \leq_\varphi w_2$ and $w_2 \leq_\varphi w_3$, then $w_1 \leq_\varphi w_3$.

(e_3) $w_1 \leq_\varphi w_2$ or $w_2 \leq_\varphi w_1$.

(e_4) If $\varphi_1 \equiv \varphi_2$, then it holds that if $w_1 \leq_{\varphi_1} w_2$, then $w_1 \leq_{\varphi_2} w_2$.

(e_5) If $w_1, w_2 \notin [\varphi]$, then $w_1 \approx_\varphi w_2$.

(e_6) If $w_1 \in [\varphi]$ and $w_2 \notin [\varphi]$, then $w_1 <_\varphi w_2$.

Note that properties e_{1-4} are identical to properties r_{1-4}, and together they imply that $\leq_\varphi$ is a total preorder on $\mathcal{U}$ that is also syntax insensitive. Thus, an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations satisfying properties e_{1-4} is *total* and *syntax insensitive* in the same sense as the one described in Section 3.1. Since we will not be considering partial assignments in this section, all $\mathcal{L}$-assignments on interpretations we will look at for enforcement will be total.

Properties e_{5-6} can be seen as analogues to revision properties r_5 and r_7, in that they regulate the effect of φ on the preorder $\leq_\varphi$, but they say something different from the revision properties. Property e_5 says that interpretations *not* satisfying φ are equally preferred, and property e_6 says that interpretations not satisfying φ are less preferred than any models of φ. Together, properties e_{5-6} imply that non-models of φ are the least plausible interpretations in $\leq_\varphi$. Properties e_5 and e_6 can be seen as duals of properties r_5 and r_7, respectively. Since we are dealing here only with total preorders, where revision properties r_5 and r_6 coincide, the enforcement property e_5 can be seen as a dual to both, i.e., we do not invoke an analogue for the revision property r_6.

To fix notation, an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations is *e-faithful* if it satisfies properties e_{5-6}. A schematic depiction of a preorder in an e-faithful assignment is given in Figure 3.7. Note that the models of $\neg\varphi$ are at the very top.

Enforcement as choice over outcomes

The next step in modeling enforcement as a choice procedure is to link up enforcement as an operation on formulas satisfying postulates E_{1-6} to plausibility relations satisfying

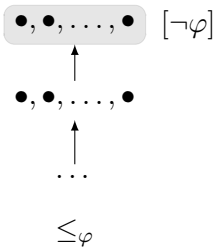


Figure 3.7: A schematic depiction of a total preorder $\leq_\varphi$ in an e-faithful assignment. As usual, bullets stand for interpretations. Models of $\neg\varphi$ (i.e., interpretations not satisfying φ) are shaded in gray.

properties e_{1-6}. This is achieved via a choice procedure, which in the case of revision and update amounts to picking the best outcomes from the models of the newly acquired information μ, in effect removing models of μ that are not optimal. However, the nature of the enforcement postulates points to a choice procedure that is in many ways different from that of revision and update: instead of removing models from μ, an enforcement operator wants to add to the models of μ: ideally, it adds all the models of φ. But if this is not possible (in case $\varphi \vee \mu$ is a tautology), some models of φ will have to be discarded. This is still an optimization-focused behavior, but the parameters under which it functions are new: using a plausibility ranking on interpretations in this setting becomes a question of not which outcomes are more readily held on to, but which are more readily given up: a small, but, as we will see, important distinction.

To characterize enforcement we introduce a new way of choosing based on a total preorder $\leq_\varphi$ on interpretations. This method uses the preorder $\leq_\varphi$ to incrementally add interpretations to $[\mu]$, until further addition becomes impossible. Thus, if $\mathcal{W}$ is a set of interpretations and $\leq$ is a preorder on interpretations, then, for $i \geq 1$, the *best-to-worst $\leq$-level i of $\mathcal{W}$*, denoted $lvl^i_\leq(\mathcal{W})$, is defined by taking:

$$lvl^1_\leq(\mathcal{W}) = \min_\leq(\mathcal{W}),$$
$$lvl^{i+1}_\leq(\mathcal{W}) = \min_\leq(\mathcal{W} \setminus (lvl^1_\leq(\mathcal{W}) \cup \cdots \cup lvl^i_\leq(\mathcal{W}))).$$

The intuition is that the elements on level i are the i^{th} best elements of $\mathcal{W}$, according to $\leq$: we intend to construct the set of models of $\varphi \triangleright \mu$ iteratively, by adding interpretations to μ in successive steps, and the $\leq$-levels of $\mathcal{W}$ will provide the order in which to do so. It is straightforward to see that the best-to-worst levels form a partition of the set $\mathcal{W}$.

Next, we must specify how to actually construct $[\varphi \triangleright \mu]$. If $\mathcal{W}$ is a set of interpretations, the *addition operator* $\mathrm{add}^i_\leq(\mathcal{W})$ is defined, for $i \geq 1$, as follows:

$$\mathrm{add}^1_\leq(\mathcal{W}) = \mathcal{W},$$
$$\mathrm{add}^i_\leq(\mathcal{W}) = \begin{cases} \mathrm{add}^{i-1}_\leq(\mathcal{W}) \cup lvl^{i-1}_\leq(\mathcal{U} \setminus \mathcal{W}), & \text{if } \mathrm{add}^{i-1}_\leq(\mathcal{W}) \cup lvl^{i-1}_\leq(\mathcal{U} \setminus \mathcal{W}) \neq \mathcal{U}, \\ \mathrm{add}^{i-1}_\leq(\mathcal{W}), & \text{otherwise.} \end{cases}$$

Intuitively, the addition operator starts from $\mathcal{W}$ and iteratively adds interpretations that are not already in $\mathcal{W}$, in the order prescribed by $\leq_\varphi$. Addition of new interpretations is controlled by an acceptance condition, saying that the new result should not be a tautology. If the acceptance condition is satisfied then the interpretations under considerations are added and the operator moves on to the next level; if not, the operator falls back to the result obtained at the previous level. The starting point guarantees that $\mathcal{W}$ is included in $\mathrm{add}^i_\leq(\mathcal{W})$, for $i \geq 0$. Note that, since $\mathcal{W}$ is finite and $\leq$ is a total preorder, this operation also reaches a fixed point, i.e., there exists an $i \in \mathcal{W}$ such that $\mathrm{add}^j_\leq(\mathcal{W}) = \mathrm{add}^i_\leq(\mathcal{W})$, for any $j > i$. Thus, if $\mathcal{W}$ is a set of interpretations and $\leq$ is a total preorder on $\mathcal{W}$, then *the fixed point of the operator* add is denoted by $\mathrm{add}^*_\leq(\mathcal{W})$.

With the notion of the addition operator in hand, we can now define a choice procedure that exploits a total preorder on interpretations to yield the result of enforcing μ with respect to φ. Thus, if $\preccurlyeq$ is a total, syntax insensitive and e-faithful assignment, the *$\preccurlyeq$-induced $\mathcal{L}$-enforcement operator* $\triangleright^\preccurlyeq$ is defined, for any propositional formulas φ and μ, as follows:

$$[\varphi \triangleright^\preccurlyeq \mu] \stackrel{\text{def}}{=} \mathrm{add}^*_{\leq_\varphi}[\mu].$$

Conversely, we want to use an $\mathcal{L}$-enforcement operator $\triangleright$ to infer a ranking over interpretations, under the assumption that choice is made using the iterative approach described above. For revision and update, we would do this using the $\mathcal{L}$-proxy of a pair of interpretations w_1 and w_2, i.e., a propositional formula $\varepsilon_{1,2}$ such that $[\varepsilon_{1,2}] = \{w_1, w_2\}$, to ask the agent which of the two outcomes it wants to hold on to. Enforcement, which, as we have seen, is a kind of dual of revision, requires a different tactic: we will use the $\mathcal{L}$-antiproxy of a pair of interpretations w_1 and w_2, i.e., a propositional formula $\varepsilon_{-1,-2}$ such that $[\varepsilon_{-1,-2}] = \mathcal{U} \setminus \{w_1, w_2\}$, to ask the agent which of the two outcomes it wants to give up. Then, if $\triangleright$ is an enforcement operator, the *$\triangleright$-revealed relation* $\leq_\varphi$ is defined by taking, for any interpretations w_1 and w_2:

$$w_1 \leq^\triangleright_\varphi w_2 \text{ if } w_2 \notin [\varphi \triangleright \varepsilon_{-1,-2}].$$

The rationale here is that w_2 is less preferred than w_1 if it is more readily given up: since the rules of enforcement say that $\varepsilon_{-1,-2} \subseteq [\varphi \triangleright \varepsilon_{-1,-2}] \subset \mathcal{U}$, interpretations w_1 and w_2 cannot both be added to $\varphi \triangleright \varepsilon_{-1,-2}$ so a choice must be as to which to give up. If w_2, rather than w_1, is given up, this indicates that w_1 is preferred to w_2; giving both of them up means that they are equally preferred.

Example 3.13: The art of diagnosis, using a preorder on outcomes

For the setting in Example 3.11, with $\varphi = a \vee b$ and $\mu = c$, we have that $[\varphi] = \{a, b, ab, ac, bc, abc\}$ and $[\mu] = \{c, ac, bc, abc\}$. Consider, first, a total, syntax independent $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that assigns to φ the preorder $\leq_\varphi$ in Figure 3.13. We obtain $[\varphi \triangleright \mu]$ by applying the addition operator add to $[\mu]$. The addition operator starts from $[\mu]$ and takes the levels of $\neg\mu$ in order, trying to add them to μ

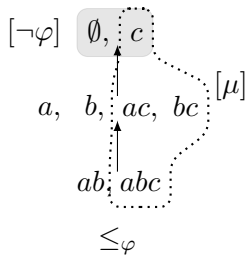


Figure 3.8: A total preorder $\leq_\varphi$ assigned to φ, with $[\varphi] = \{a, b, ab, ac, bc, abc\}$, by a total, syntax insensitive and e-faithful assignment. Note that the interpretations not satisfying φ, i.e., models of $[\neg\varphi]$, are the least preferred outcomes in this preorder and are in the gray region. The new information is μ, with $[\mu] = \{c, ac, bc, abc\}$. Enforcing μ with respect to φ involves adding interpretations to $[\mu]$ that are not already in $[\mu]$ in the order prescribed by $\leq_\varphi$, unless a tautology is created.

while avoiding the creation of a tautology. The operation is successful for the first and second levels, after which a fixed point is reached. The result is:

$$\begin{aligned}[\varphi \triangleright \mu] &= \mathrm{add}^*_{\leq_\varphi}[\mu] \\ &= ([\mu] \cup \{ab\}) \cup \{a, b\} \\ &= \{a, b, c, ab, ac, bc, abc\}.\end{aligned}$$

Converting the result back into a propositional formula, we obtain that $\varphi \triangleright \mu \equiv a \vee b \vee c$.

Conversely, the ranking between two outcomes, say a and ab, relative to the prior information φ, with $[\varphi] = \{a, b, ab, ac, bc, abc\}$, can be elicited by asking the agent to enforce the new information $\varepsilon_{-1,-2}$, with $[\varepsilon_{-1,-2}] = \{\emptyset, b, c, ac, bc, abc\}$. Supposing the result is $[\varphi \triangleright \varepsilon_{-1,-2}] = [\varepsilon_{-1,-2}] \cup \{ab\}$, we can conclude that, according to the $\triangleright$-revealed ranking, it holds that $ab <^{\triangleright}_{\varphi} a$. This is consistent with $\leq_\varphi$ as depicted in Figure 3.8.

The test of our construction, of course, is whether postulates E_{1-6}, properties e_{1-6} and the choice procedure formalized by the addition operator add work together to describe a single belief change mechanism. The validation comes in the form of a representation theorem, which shows that these notions cohere with each other. Before introducing the result, though, we need to explain what it means for an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations to represent an enforcement operator. Thus, if $\triangleright$ is an $\mathcal{L}$-enforcement operator and $\preccurlyeq$ is an $\mathcal{L}$-assignment on interpretations, then $\preccurlyeq$ *represents* $\triangleright$ (alternatively, $\triangleright$ *is represented by* $\preccurlyeq$) if, for any propositional formulas φ and μ, it holds that $[\varphi \triangleright \mu] = \mathrm{add}^*_{\leq_\varphi}[\mu]$. We can now introduce the main representation theorem of this section.

Theorem 3.9

An $\mathcal{L}$-enforcement operator $\triangleright$ satisfies postulates E_{1-6} if and only if there exists an $\mathcal{L}$-assignment $\preccurlyeq$ that satisfies properties e_{1-6} (i.e., that is total, syntax insensitive and e-faithful) and that represents the operator $\triangleright$.

Proof

("$\Leftarrow$") Note that a preorder $\leq_\varphi$ that satisfies properties e_{1-6} can be seen as a preorder $\leq_{\neg\varphi}$, i.e., a preorder depending on $\neg\varphi$, by taking $w_1 \leq_{\neg\varphi} w_2$ if $w_2 \leq_\varphi w_1$, i.e., by turning $\leq_\varphi$ upside down. In this case, $\leq_{\neg\varphi}$ satisfies the revision properties r_{1-7}. Note that in this setting we have that the complement of $\mathrm{add}^*_{\leq_\varphi}[\mu]$ consists of the minimal models of $\neg\mu$ in the preorder $\neg\varphi$ as just defined. We can now see that the upside down assignment corresponds to a total, syntax insensitive r-faithful assignment, which corresponds to a revision operator $\circ$ that satisfies postulates R_{1-6}. Furthermore, we get that $[\varphi \triangleright \mu] = \mathcal{U} \setminus (\neg\varphi \circ \neg\mu)$, which, by Proposition 3.6, implies that $\triangleright$ satisfies postulates E_{1-6}.

("$\Rightarrow$") The $\triangleright$-revealed assignment is the assignment we are looking for, and it is straightforward to check that it satisfies properties e_{1-6} and that it represents the operator $\triangleright$.

A quick note is in order on previous results. Existing work on propositional enforcement [Haret et al., 2018c] has used partial orders on sets of interpretations (or, alternatively, on formulas) to represent enforcement operators, but a nice representation in terms of preorders on interpretations themselves, à la Theorem 3.5, was left open. Here we filled this gap. Note, also, that a representation in terms of preorder on interpretations can be obtained in a more naive way, by using Equations 3.1 and 3.2 and Theorem 3.1. Thus, given an enforcement operator $\triangleright$ satisfying postulates E_{1-6}, we can immediately infer that there exists an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that satisfies properties r_{1-5} and r_7 such that, for any propositional formulas φ and μ, the following holds:

$$[\varphi \triangleright \mu] = [\neg(\neg\varphi \circ^{\triangleright} \neg\mu)] = \mathcal{U} \setminus \min_{\leq_{\neg\varphi}}[\neg\mu].$$

In other words, we can use an assignment representing the $\triangleright$-induced $\mathcal{L}$-revision operator $\circ^{\triangleright}$ to represent $\triangleright$. However, this expression is not very informative and, as we have shown, unnecessarily circuitous.

Distance-based enforcement operators

Theorem 3.9 offers some insight into how to construct concrete enforcement operators: find a way to generate e-faithful assignments, i.e., preorders on interpretations in which the top elements are the non-models of φ. It turns out that the tried and tested methods of quasi-distances between interpretations and aggregation functions used in revision and update work here as well, with minimal adjustment.

We bring up aggregation functions only to settle straight away that the only aggregation function we will use in this section is the min aggregation function: we will be using it, however, in a slightly different way than in revision. For the next definition, recall that the size of the set A is assumed to be m. If d is a quasi-distance between interpretations, φ is a propositional formula and w is an interpretation, then the $(d,\ m-\min)$-*induced distance* $d^{m-\min}(\varphi, w)$ *between* φ *and* w is defined as:

$$d^{m-\min}(\varphi, w) = m - \min(d(v, w))_{v\in[\varphi]}.$$

Intuitively, $d^{m-\min}(\varphi, w)$ can be thought of as the inverse of the more familiar notion $d^{\min}(\varphi, w)$ (see Sections 2.3 or 3.1): a penalty is introduced for w the closer it is to the models of φ, such that the interpretations closest to φ end up receiving the highest score and the interpretations furthest to φ receive the lowest score, i.e., φ is such that it is better to be far away from it than close to it.

Following up, the $(d,\ m-\min)$-*induced ranking* $\leq_\varphi^{d,\ m-\min}$ is defined as:

$$w_1 \leq_\varphi^{d,\ m-\min} w_2 \text{ if } d^{m-\min}(\neg\varphi, w_1) \leq d^{m-\min}(\neg\varphi, w_2).$$

What we are saying, in effect, is that w_1 is preferred to w_2 relative to φ, according to $\leq_\varphi^{d,\ m-\min}$ if w_1 is farther away from $\neg\varphi$ than w_2. If d is a quasi-distance, the $(d,\ m-\min)$-*induced assignment* $\preccurlyeq^{d,\ m-\min}$ is obtained by taking $\preccurlyeq^{d,\ m-\min}(\varphi) = \leq_\varphi^{d,\ m-\min}$, for any propositional formula φ. In the same vein, the $(d,\ m-\min)$-*induced* $\mathcal{L}$*-enforcement operator* $\triangleright^{d,\ m-\min}$ is the operator induced by the assignment $\preccurlyeq^{d,\ m-\min}$. This allows us to generate total, syntax insensitive e-faithful assignments.

Proposition 3.7

If d is a quasi-distance between interpretations and φ is a propositional formula, the $(d,\ m-\min)$-induced ranking $\leq_\varphi^{d,\ m-\min}$ satisfies properties e_{1-6}, i.e., $\leq_\varphi^{d,\ m-\min}$ is a total preorder on interpretations that is syntax insensitive and makes the models of $\neg\varphi$ the $\leq_\varphi^{d,\ m-\min}$-maximal elements.

Proof

It is straightforward to see that $\leq_\varphi^{d,\ m-\min}$ is a total preorder on interpretations, i.e., that $\leq_\varphi^{d,\ \min}$ satisfies properties e_{1-3}. Since the definition of $\leq_\varphi^{d,\ m-\min}$ depends only on the interpretations, $\leq_\varphi^{d,\ m-\min}$ also satisfies property e_4. Finally, it holds that: $d^{m-\min}(\neg\varphi, w) = m - \min(d(v, w))_{v\in[\neg\varphi]} = m$ if and only if $w \in [\neg\varphi]$, which implies that models of $\neg\varphi$ are the $\leq_\varphi^{d,\ m-\min}$-maximal elements in $\leq_\varphi^{d,\ m-\min}$, i.e., $\leq_\varphi^{d,\ m-\min}$ satisfies properties e_5 and e_6, for any propositional formula φ.

Proposition 3.7 implies that the $(d,\ m-\min)$-induced assignment $\preccurlyeq^{d,\ m-\min}$ is total, e-faithful and syntax insensitive, which, by Theorem 3.9, implies that the $(d,\ m-\min)$-induced $\mathcal{L}$-enforcement operator $\triangleright^{d,\ \min}$ satisfies postulates E_{1-6}.

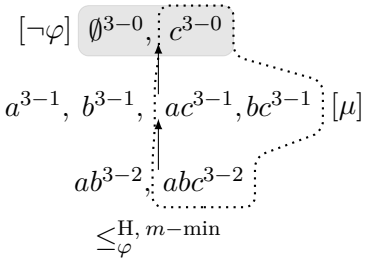


Figure 3.9: The preorder $\leq_\varphi^{\mathrm{H},\, m-\min}$ assigned to φ, with $[\varphi] = \{a, b, ab, ac, bc, abc\}$, by the $\preccurlyeq^{\mathrm{H},\, m-\min}$ assignment. The superscripts denote the distance to $\neg\varphi$ subtracted from the number of atoms in A, which in this case is 3. Note that the interpretations not satisfying φ, i.e., models of $[\neg\varphi]$, get a score of 3, which makes them the least preferred outcomes.

Corollary 3.5

If d is a quasi-distance between interpretations, the $(d, m{-}\min)$-induced $\mathcal{L}$-enforcement operator $\triangleright^{d,\, m-\min}$ satisfies postulates E_{1-6}.

As examples of concrete distances we can use the Hamming distance d_{H} and the drastic distance d_{D}, giving rise to the $\mathcal{L}$-enforcement operators $\triangleright^{\mathrm{H},\, m-\min}$ and $\triangleright^{\mathrm{D},\, m-\min}$.

Example 3.14: The art of diagnosis, using distances

For the setting in Example 3.11, with $A = \{a, b, c\}$, $\varphi = a \vee b$ and $\mu = c$, the $\mathcal{L}$-assignment $\preccurlyeq^{\mathrm{H},\, m-\min}$ generates the preorder $\leq_\varphi^{\mathrm{H},\, m-\min}$ in Figure 3.9. Note that $\leq_\varphi^{\mathrm{H},\, m-\min}$ is the same as the preorder $\leq_\varphi$ in Figure 3.8. We obtain, as before, that $[\varphi \triangleright \mu] = \mathrm{add}^*_{\leq_\varphi}[\mu] = \{a, b, c, ab, ac, bc, abc\}$.

3.4 Merging

If agents deliberate with respect to a small number of independent alternatives, as is the case in a typical election, aggregation of different viewpoints is well understood due to existing research in the field of social choice [Zwicker, 2016, Baumeister and Rothe, 2016]. But if agents have to decide on multiple interconnected issues at the same time, then the number of possible alternatives can grow too large to expect agents to have explicit preferences over the whole set. We have encountered this kind of scenario in Example 1.5, where we have been introduced to four Academy members trying to decide who will be the nominees in this year's *Best Director* category.

Example 3.15: #OscarsSoFossilized

In Example 1.5 the names being circulated are Alma Har'el, Bong Joon Ho and Céline Sciamma, represented by propositional variables a, b and c, respectively. The decision as to who will be the nominees is left up to four Academy members. Each of the four members has their own opinion about who should be nominated, represented by propositional formulas $\varphi_1 = a \wedge b$, $\varphi_2 = a \wedge (b \vee c)$, $\varphi_3 = \neg a \wedge b \wedge \neg c$. and $\varphi_4 = \neg a \wedge \neg b \wedge c$. The final lineup should consist of only two people, i.e., the individual opinions should be aggregated subject to the constraint $\mu = (a \wedge b \wedge \neg c) \vee (a \wedge \neg b \wedge c) \vee (\neg a \wedge b \wedge c)$. These opinions are collectively inconsistent, but none of them weighs more than the others.

In Example 3.15, a standard social choice procedure would require the Academy members to provide a ranking of all the possible lists of two nominees, or, as we code them here, of the outcomes ab, bc, ac. Though this would not be too difficult for this example, the cognitive burden on the agents will certainly become too big if the number of possibilities or the size of the lineup grew even slightly. This is certain to be the case even for the Oscars: Example 3.15 is only a toy example, since the real world list of *Best Director* nominees is usually made up of five people, and the list of possible nominees is much larger. In the real world scenario, asking Academy members to rank all possible combinations of five directors is clearly unfeasible.

This problem, known more generally as *combinatorial voting* [Lang and Xia, 2016], acquires a knowledge representation dimension as agents need compact ways to express their positions over a large domain, and automatizable procedures to perform reasoning with such positions. Merging, in this context, can prove useful, as it provides a versatile framework in which different agents can combine their positions on a fixed set of issues, expressed as propositional formulas, into a collective perspective, expressed, likewise, as a propositional formula [Konieczny and Pino Pérez, 2002, Konieczny and Pino Pérez, 2011].

In Example 3.15 a merging operator should combine the information provided by the four Academy members while making sure that the cardinality constraint μ is satisfied. What the theory of belief merging offers is a core set of postulates to assess the rationality of any merging operator, and a range of concrete operators tailored according to these principles. Seeing merging operators as a type of collective decision procedure is a natural interpretation of the process: the propositional atoms in the alphabet can be taken to encode issues that are deliberated upon, while truth-value assignments to atoms, i.e., the interpretations or outcomes, encode combinations of issues that could make it into the final result, and over which agents can have preferences. The propositional formulas submitted by agents represent the way in which issues are interconnected in the agents' preferences, and the result is a set of "winning" interpretations, representable as a propositional formula, that respect the integrity constraint of the merging process.

Postulates

An $\mathcal{L}^n$-*merging operator* Δ is a function $\Delta\colon \mathcal{L}^n \to \mathcal{L}$, taking as input a propositional profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$ and a propositional formula μ, and returning a propositional formula, denoted by $\Delta_\mu(\vec{\varphi})$. In the context of merging the formula μ, which we normally call the new information, is called an *integrity constraint*. Merging is a multi-agent operation, in the sense that the formulas in a profile $\vec{\varphi}$ originate with different agents, usually gathered in the set $N = \{1, \ldots, n\}$, with formula φ_i corresponding to agent i.

The following postulates are typically taken to provide a core set of rationality constraints any merging operator Δ is expected to satisfy. They are expected to hold for any propositional profiles $\vec{\varphi}$, $\vec{\varphi}_1$, $\vec{\varphi}_2$, formulas φ_1, φ_2 and constraints μ, μ_1 and μ_2:

(M_0) $\Delta_\mu(\vec{\varphi}) \models \mu$.

(M_1) If μ is consistent, then $\Delta_\mu(\vec{\varphi})$ is consistent.

(M_2) If $\bigwedge \vec{\varphi} \wedge \mu$ is consistent, then $\Delta_\mu(\vec{\varphi}) \equiv \bigwedge \vec{\varphi} \wedge \mu$.

(M_3) If $\vec{\varphi}_1 \equiv \vec{\varphi}_2$ and $\mu_1 \equiv \mu_2$, then $\Delta_{\mu_1}(\vec{\varphi}_1) \equiv \Delta_{\mu_1}(\vec{\varphi}_2)$.

(M_4) If $\varphi_1 \models \mu$ and $\varphi_2 \models \mu$, then $\Delta_\mu(\varphi_1, \varphi_2) \wedge \varphi_1$ is consistent if and only if $\Delta_\mu(\varphi_1, \varphi_2) \wedge \varphi_2$ is consistent.

(M_5) $\Delta_\mu(\vec{\varphi}_1) \wedge \Delta_\mu(\vec{\varphi}_2) \models \Delta_\mu(\vec{\varphi}_1 + \vec{\varphi}_2)$.

(M_6) If $\Delta_\mu(\vec{\varphi}_1) \wedge \Delta_\mu(\vec{\varphi}_2)$ is consistent, then $\Delta_\mu(\vec{\varphi}_1 + \vec{\varphi}_2) \models \Delta_\mu(\vec{\varphi}_1) \wedge \Delta_\mu(\vec{\varphi}_2)$.

(M_7) $\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2 \models \Delta_{\mu_1 \wedge \mu_2}(\vec{\varphi})$.

(M_8) If $\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$ is consistent, then $\Delta_{\mu_1 \wedge \mu_2}(\vec{\varphi}) \models \Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$.

Postulate M_0 says that the merging result $\Delta_\mu(\vec{\varphi})$ should satisfy the constraint μ. Postulate M_1 says that the result should be consistent if μ is consistent. Postulate M_2 requires that if there is any agreement between the formulas in $\vec{\varphi}$ and μ, then the merged result is nothing more than the agreed upon outcomes. Postulate M_3 says that the result should be insensitive to the syntax of the formulas involved. Postulate M_4 stipulates that merging two formulas φ_1 and φ_2 should be fair, in the sense that if the result contains outcomes consistent with one of the formulas, it should contain results consistent with the other as well. Postulates M_{5-6} say that the result should include outcomes that are unanimously accepted across subprofiles. Postulates M_{7-8} say that the result and coherent when varying the constraint.

Though postulates M_{0-8} are referred to here using our custom naming convention, in the literature they are more commonly known as the *IC*-postulates [Konieczny and Pino Pérez, 2002, Konieczny and Pino Pérez, 2011], where '*IC*' stands for *integrity constraint*

and indicates that merging is done within the purview of the condition μ, which must be satisfied by the merging result $\Delta_\mu(\vec{\varphi})$.

Note that, insofar as a profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ is identified with the single propositional formula $\bigwedge_{\varphi_i \in \vec{\varphi}} \varphi_i$, then postulates M_{0-3} and M_{7-8} correspond to revision postulates R_{1-4} and R_{5-6}, respectively, where $\bigwedge \vec{\varphi}$ is the prior belief and μ is the newly acquired information. Thus, another way of looking at a merging operator Δ is to see it as a revision operator that needs to satisfy some additional properties, besides the standard ones presented in Section 3.1. These properties are postulates M_{4-6}, and what they add is the notion that the formulas that go into the prior belief (or rather, their models) should carry equal weight in the change process. This corresponds to the idea that merging is a public, or social operation, whose participants should be treated fairly. Consequently, postulates M_{0-8} are best understood as axiomatizing a decision procedure based on the aggregation of information coming from different sources, i.e., the formulas in $\vec{\varphi}$.

Example 3.16: Possible Oscar nominees

For the merging scenario in Example 3.15 the profile is $\vec{\varphi} = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$, where $\varphi_1 = a \wedge b$, $\varphi_2 = a \wedge (b \vee c)$, $\varphi_3 = \neg a \wedge b \wedge \neg c$ and $\varphi_4 = \neg a \wedge \neg b \wedge c$. Examples 1.5 and 3.15 provide the meaning for these formulas. The constraint is represented by the propositional formula $\mu = (a \wedge b \wedge \neg c) \vee (a \wedge \neg b \wedge c) \vee (\neg a \wedge b \wedge c)$, with $[\mu] = \{ab, bc, ac\}$.

Suppose that Δ is a merging operator that satisfies postulates M_{0-8} and $\Delta_\mu(\vec{\varphi}) \equiv (a \wedge b) \vee (b \wedge c)$, i.e., $[\Delta_\mu(\vec{\varphi})] = \{ab, bc\}$. This result is in line with postulate M_1 (i.e., it is consistent) and with postulate M_0 (i.e., implies μ). The models of $[\Delta_\mu(\vec{\varphi})]$ encode the makeup of the list of nominees, e.g., ab means that Alma Har'el and Bong Joon Ho (but not Céline Sciamma) will be nominated.

As with revision and the other belief change operations we have looked at in this chapter, the positions of the agents could be beliefs, intentions or simply combinations of issues that agents find desirable, and would like to see in the result. The consistent insistence on insensitivity to syntax means that the propositional formulas that stand for the agents' positions are, in a sense, just window dressing for their models. This is another way of saying that belief change operations are interested more in the underlying issues rather than in how they are expressed. Of course, the representation may matter for cognitive or computational purposes, but for the belief change operators we consider here the formulas are just compact representations of sets of outcomes the agents are interested in. Thus, whether or not the formulas encode (actual) beliefs is not of immediate crucial importance to a merging operator: postulates M_{0-8} are neutral with respect to the cognitive attitude being expressed.

That being said, the exact meaning of the formulas *will* matter if we want to be more specific about the type of information aggregation a belief merging operator performs. Thus, there is a significant difference between aggregating bits of information the agents

believe are true, when there is, actually, a true state of the world, versus aggregating sets of issues the agent want to see obtain, in which case there might not be a true answer at all. In the former case the purpose of a merging operator is to track the truth, whereas in the latter case the purpose of a merging operator is to be fair towards the participants. Correspondingly, the criteria a merging operator is expected to satisfy will be different depending on the kind of task it is used for: a truth-tracking operator will be expected to be accurate, whereas a fair merging operator will be expected to be impartial towards the agents, strategyproof or proportional.

In this work we are interested in merging more as a tool for collective decision making than as a way of aggregating information about the world, and will therefore focus on the fairness aspects of merging. Work on the truth-tracking abilities of merging exists [Everaere et al., 2010b], but is outside the scope of the current work.

Preferences over outcomes

In the context of merging, preference orders are ushered in through an $\mathcal{L}^n$-*assignment* $\preccurlyeq$ *on interpretations*, which is a function $\preccurlyeq\colon \mathcal{L}^n \to 2^{\mathcal{U}\times\mathcal{U}}$ that maps $\mathcal{L}$-profiles to preference orders on interpretations. In Sections 3.1 and 3.2 we have modeled both total and partial preorders, but here we will work mainly with total preorders. Since we want to pursue the parallel between merging and a collective decision process, the relation $\leq_{\vec{\varphi}}$ assigned to a profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$ by an $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations can be thought of as the collective ranking, obtained by aggregating the preferences of each agent in $N = \{1, \ldots, n\}$. In this context, the preferences of the agents themselves are given by $\leq_{(\varphi_i)}$, i.e., by the preference associated with the profile (φ_i) having only φ_i as element. In the interest of readability, we will generally write simply $\leq_{\varphi_i}$ instead of $\leq_{(\varphi_i)}$, when only one agent is involves.

Under the assumption of an $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations, we have both that the individual preference orders $\leq_{\varphi_i}$, for $i \in N$, as well as the collective preference $\leq_{\vec{\varphi}}$, exist. The purpose of merging, however, is to make sure not only that the individual and collective preference orders exist, but that they also have desirable properties, i.e., that the collective preference order can be seen to aggregate, in a fair and reasonable way, the information provided by the individual preference orders. This is ensured by writing down desirable properties of an $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations. The properties an $\mathcal{L}^n$-assignment is expected to satisfy include some familiar properties, but also some new ones. Recall that two propositional profiles $\vec{\varphi}_1$ and $\vec{\varphi}_2$ are equivalent, written $\vec{\varphi}_1 \equiv \vec{\varphi}_2$, if there is a bijection $f\colon \vec{\varphi}_1 \to \vec{\varphi}_2$ such that $f(\varphi_i) \equiv \varphi_i$, for any $\varphi_i \in \vec{\varphi}_1$. The properties are expected to apply for any propositional profiles $\vec{\varphi}$, $\vec{\varphi}_1$, $\vec{\varphi}_2$, propositional formulas φ_1 and φ_2, and interpretations w, w_1 and w_2:

(m_1) $w \leq_{\vec{\varphi}} w$.

(m_2) If $w_1 \leq_{\vec{\varphi}} w_2$ and $w_2 \leq_{\vec{\varphi}} w_3$, then $w_1 \leq_{\vec{\varphi}} w_3$.

(m_3) $w_1 \leq_{\vec{\varphi}} w_2$ or $w_2 \leq_{\vec{\varphi}} w_1$.

(m_4) If $\vec{\varphi}_1 \equiv \vec{\varphi}_2$, then $\leq_{\vec{\varphi}_1} = \leq_{\vec{\varphi}_2}$.

(m_5) If $w_1, w_2 \in [\vec{\varphi}]$, then $w_1 \approx_{\vec{\varphi}} w_2$.

(m_6) If $w_1 \in [\vec{\varphi}]$ and $w_2 \notin [\vec{\varphi}]$, then $w_1 <_{\vec{\varphi}} w_2$.

(m_7) If φ_1 and φ_2 are consistent and $w_1 \in [\varphi]$, then there exists $w_2 \in [\varphi_2]$ such that $w_2 \leq_{(\varphi_1,\varphi_2)} w_1$.

(m_8) If $w_1 \leq_{\vec{\varphi}_1} w_2$ and $w_1 \leq_{\vec{\varphi}_2} w_2$, then $w_1 \leq_{\vec{\varphi}_1+\vec{\varphi}_2} w_2$.

(m_9) If $w_1 \leq_{\vec{\varphi}_1} w_2$ and $w_1 <_{\vec{\varphi}_2} w_2$, then $w_1 <_{\vec{\varphi}_1+\vec{\varphi}_2} w_2$.

Properties m_{1-3} imply that $\leq_{\vec{\varphi}}$ is a total preorder on interpretations, and are identical to revision properties r_{1-3}. Property m_4 expresses syntax insensitivity in the context of merging. Properties m_{5-6} say that models of a profile $\vec{\varphi}$ are the uniquely minimal elements in $\leq_{\vec{\varphi}}$, and are equivalent to revision properties r_5 and r_7, respectively. Property m_7 says that models of two formulas φ_1 and φ_2 should be treated equally when merging φ_1 and φ_2, in the sense that it should not be the case that some model of φ_1 gets chosen while all models of φ_2 are left out (assuming it is possible to choose models from both φ_1 and φ_2). Property m_8 says that if w_1 is considered at least as good as w_2 according to both profiles $\vec{\varphi}_1$ and $\vec{\varphi}_2$, then w_1 is also considered at least as good as w_2 according to the profile $\vec{\varphi}_1 + \vec{\varphi}_2$, obtained by concatenating $\vec{\varphi}_1$ and $\vec{\varphi}_2$; in other words, agreement about w_1 and w_2 carries over to the aggregated result. Property m_9 says that if if w_1 is considered at least as good as w_2 according to both profiles $\vec{\varphi}_1$ and $\vec{\varphi}_2$ and, in addition, w_1 is considered strictly better than w_2 according to at least one of the profiles $\vec{\varphi}_1$ and $\vec{\varphi}_2$, then w_1 is considered strictly better than w_2 according to the profile $\vec{\varphi}_1 + \vec{\varphi}_2$, obtained by concatenating $\vec{\varphi}_1$ and $\vec{\varphi}_2$.

An $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations is *total* if it satisfies properties m_{1-3}, *syntax insensitive* if it satisfies propery m_4 and *m-faithful* if it satisfies properties m_{5-9}. A total, syntax insensitive and m-faithful assignment corresponds to what is more usually called a *syncretic assignment* [Konieczny and Pino Pérez, 2002, Konieczny and Pino Pérez, 2011]. A schematic depiction of such an assignment is offered in Figure 3.10.

Merging as social choice over outcomes

Apart from the multi-agent flavour given by the extra postulates, merging can be formalized as a bona-fide belief change operator along the same lines as revision, update and enforcement. This means using the preference information afforded by an $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations to obtain the result of merging formulas in a profile and, conversely, using the merging result to infer the underlying preference relation. Recall, for this, that the $\mathcal{L}$-proxy of a pair $\{w_1, w_2\}$ of interpretations is a propositional formula $\varepsilon_{1,2}$ such that $[\varepsilon_{1,2}] = \{w_1, w_2\}$.

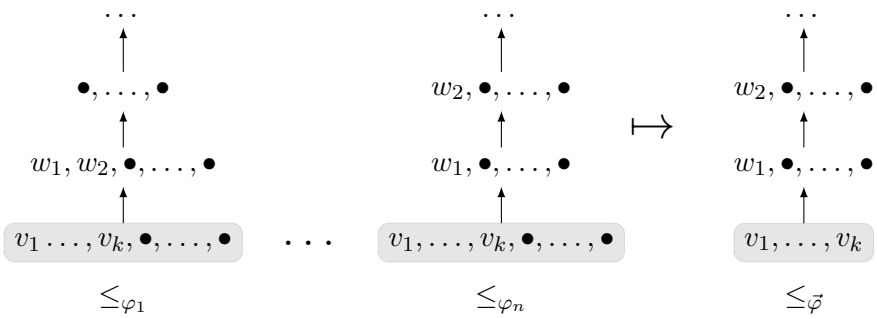


Figure 3.10: A schematic depiction of preorders $\leq_{\varphi_i}$, $\leq_{\varphi_n}$ and $\leq_{\vec{\varphi}}$ in an m-faithful assignment. Bullets stand for interpretations. Models of φ_i, for $1 \leq i \leq n$, and of $\vec{\varphi}$ are shaded in light gray. Note that all agents in $\vec{\varphi}$ agree that w_1 is at least as good as w_2 and, in accordance with m_8, w_1 is at least as good as w_2 in $\leq_{\vec{\varphi}}$. What is more, in $\leq_{\varphi_n}$ it holds that w_1 is strictly preferred to w_2, and in accordance with m_9, w_1 is strictly better than w_2 in $\leq_{\vec{\varphi}}$. Note, also, that $v_1, \ldots, v_k$ are models of every formula in $\vec{\varphi}$, and are among the minimal elements in each preorder $\leq_{\varphi_i}$, for $1 \leq i \leq n$. In addition, the models of $\vec{\varphi}$, i.e., $v_1, \ldots, v_k$, are the uniquely minimal elements in $\leq_{\vec{\varphi}}$.

Thus, if $\preccurlyeq$ is an $\mathcal{L}^n$-assignment on interpretations, the *$\preccurlyeq$-induced $\mathcal{L}^n$-merging operator* $\Delta^{\preccurlyeq}$ is defined, for any profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ and constraint μ, as:

$$[\Delta_\mu(\vec{\varphi})] \stackrel{\text{def}}{=} \min_{\leq_{\vec{\varphi}}}[\mu].$$

Conversely, if Δ is a merging operator, then the *Δ-revealed relation $\leq^{\Delta}_{\vec{\varphi}}$ on interpretations* is defined, for any propositional profile $\vec{\varphi}$ and interpretations w_1 and w_2, as follows:

$$w_1 \leq^{\Delta}_{\vec{\varphi}} w_2 \text{ if } w_1 \in [\Delta_{\varepsilon_{1,2}}(\vec{\varphi})].$$

Predictably, the *Δ-revealed $\mathcal{L}^n$-assignment $\preccurlyeq^{\Delta}$ on interpretations* is defined, for any propositional profile $\vec{\varphi}$, as $\preccurlyeq^{\Delta}(\vec{\varphi}) = \leq_{\vec{\varphi}}$. If Δ is an $\mathcal{L}^n$-merging operator and $\preccurlyeq$ is an $\mathcal{L}^n$-assignment on interpretations, then $\preccurlyeq$ *represents* Δ (and, alternatively, Δ *is represented by* $\preccurlyeq$), if, for any $\mathcal{L}$-profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ and constraint μ, it holds that $[\Delta_\mu(\vec{\varphi})] = \min_{\leq_{\vec{\varphi}}}[\mu]$. The following classical representation theorem shows that postulates M_{0-8} and properties m_{1-9} fit together into a single choice procedure.

Theorem 3.10 ([Konieczny and Pino Pérez, 2011])

An $\mathcal{L}^n$-merging operator Δ satisfies postulates M_{0-8} if and only if there exists an $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations that satisfies properties m_{1-9} (i.e., is total, syntax insensitive and m-faithful) and that represents the operator Δ.

As for revision, postulate M_2 enjoys a one-to-one correspondence with properties m_{5-6} and can be separated from the rest of the postulates, though there is no pressing need to do so for merging operators, since we will not consider alternatives to it.

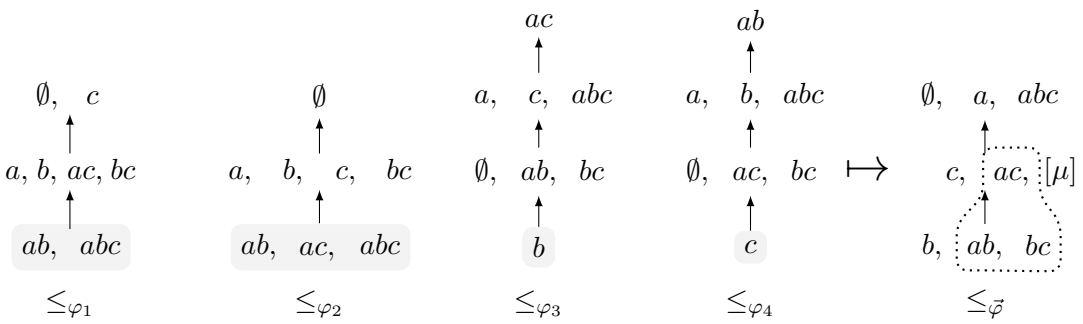


Figure 3.11: Total preorders $\leq_{\varphi_i}$, for $i \in \{1,2,3,4\}$, corresponding to the opinions of the four Academy members, as well as a preorder $\leq_{\vec{\varphi}}$, corresponding to the profile $\vec{\varphi}$. As expected, an arrow from w_1 to w_2 means that w_1 is strictly better than w_2 in the corresponding preorder, while equally preferred interpretations are separated by a comma. Note that models of φ_i are on the bottom, in $\leq_{\varphi_i}$, i.e., are the most preferred outcomes according to agent i. The models of the result are the most preferred models of the constraint μ (depicted here in the area bordered by the dotted line) in the collective preorder $\leq_{\vec{\varphi}}$.

Theorem 3.10 validates the choice perspective as applied to merging operators. According to it a merging operator that satisfies postulates M_{0-8} can be seen as a social choice function as described in Section 2.4, with a preference profile, an aggregation rule and a set of winner. The preference profile is $(\leq_{\varphi_i})_{1\leq i\leq n}$, i.e., it is made up of the individual preference orders of every agent in $\vec{\varphi}$, guaranteed to exist in the assignment that represents Δ. The merging operator is the aggregation function, and the set of winners are the models of $\Delta_\mu(\vec{\varphi})$. In fact, under the assumption of an $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations, the merging operator Δ is even a social welfare function, since the result is actually a preorder, i.e., the preorder $\leq_{\vec{\varphi}}$ associated to $\vec{\varphi}$.

Example 3.17: #OscarsSoFossilized, with preorders

For the setting in Example 3.15, the profile is $\vec{\varphi} = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$, with $\varphi_1 = a \wedge b$, $\varphi_2 = a \wedge (b \vee c)$, $\varphi_3 = \neg a \wedge b \wedge \neg c$. and $\varphi_4 = \neg a \wedge \neg b \wedge c$. The constraint is μ, with $\mu = (a \wedge b \wedge \neg c) \vee (a \wedge \neg b \wedge c) \vee (\neg a \wedge b \wedge c)$. Consider, first, a total, syntax independent and m-faithful $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations that assigns to $\vec{\varphi}$ and to the formulas in $\vec{\varphi}$ the preorders in Figure 3.11. Note that the assignment slice we have presented here is in agreement with properties m_{1-9}. According to this assignment we obtain that $[\Delta^{\preccurlyeq}_{\mu}(\vec{\varphi})] = \min_{\leq_{\vec{\varphi}}}[\mu] = \{ab, bc\}$.

Conversely, take a merging operator Δ such that $[\Delta_{\varepsilon_{ab,bc}}(\vec{\varphi})] = \{ab, bc\}$. According to the Δ-revealed assignemnt, we would infer that $ab \approx^{\Delta}_{\vec{\varphi}} bc$, which is in accordance with $\leq_{\vec{\varphi}}$ as depicted in Figure 3.11.

Distance-based merging operators

Standard ways of constructing merging operators that satisfy postulates M_{0-8} are based on the idea of finding outcomes that minimize overall distance to the profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$, and rely on ingredients that we have encountered before. The first ingredient is a distance function d on interpretations, i.e., a function $d\colon \mathcal{U}\times\mathcal{U}\to\mathbb{R}_{\geq 0}$ that satisfies properties D_{1-3} in Section 2.3. Note that, in contrast to revision, update and enforcement, which only require d to be a quasi-distance, merging requires d to be a distance. The distance d is then used to generate, for any propositional formula φ and interpretation w, the $(d,\ \min)$-induced distance $d^{\min}(\varphi, w)$ from a formula φ to an interpretation w, defined, as usual, as:

$$d^{\min}(\varphi, w) \stackrel{\text{def}}{=} \min(d(v, w))_{v\in[\varphi]}.$$

Following the custom established for the previous belief change operators, the min aggregation function used in the definition of $d^{\min}(\varphi, w)$ would count as a second parameter in the notation for the anticipated induced merging operator: however, since the merging operators we will look at in this work do not rely on any other aggregation functions at this step, we will not count it as a distinct modeling choice and omit it from the list of parameters passed on to the belief change function. Thus, in the context of merging only, we will write $d(\varphi, w)$ instead of $d^{\min}(\varphi, w)$.

Based on this notion, we can introduce the d-induced ranking on interpretations, defined, for any propositional formula φ and interpretations w_1 and w_2, as:

$$w_1 \leq_\varphi^d w_2 \text{ if } d(\varphi, w_1) \leq d(\varphi, w_2).$$

Merging does, nonetheless, appeal to ann aggregation function $\oplus$ as a second ingredient, and $\oplus$ is expected to satisfy properties Ag_{1-3} in Section 2.3. Thus, if d is a distance between interpretations, $\oplus$ is an aggregation function that satisfies properties Ag_{1-3}, $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$ is an $\mathcal{L}$-profile and w is an interpretation, the $(d,\ \oplus)$-*induced distance* $d^{\oplus}(\vec{\varphi}, w)$ *from* $\vec{\varphi}$ *to* w is defined as:

$$d^{\oplus}(\vec{\varphi}, w) \stackrel{\text{def}}{=} \oplus(d(\varphi_i, w))_{1\leq i\leq n}.$$

Consequently, the $(d,\ \oplus)$-*induced ranking on interpretations* is defined, for any profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$ and interpretations w_1 and w_2, as:

$$w_1 \leq_\varphi^{d,\,\oplus} w_2 \text{ if } d^{\oplus}(\vec{\varphi}, w_1) \leq d^{\oplus}(\varphi, w_2).$$

Note that this definition bears a strong resemblance to the definition used for defining the distance from a single formula to an interpretation in Section 3.1, in that it uses the two parameters of a distance and an aggregation function. But, as mentioned above, the aggregation function here stands for an extra aggregation step, such that if we were to follow the overall notational convention we would have to use three parameters (one distance function and two aggregation functions). Since one aggregation function is assumed to be fixed, however, we omit writing it explicitly. Note, as well, that if φ is a propositional formula, then $\leq_{(\varphi)}=\leq_\varphi$.

	d_{H}	$[\varphi_1]$ $\{ab, abc\}$	$[\varphi_2]$ $\{ab, ac, abc\}$	$[\varphi_3]$ $\{b\}$	$[\varphi_4]$ $\{c\}$	leximax	leximin	sum
	$\emptyset$	2	2	1	1	$(2,2,1,1)$	$(1,1,2,2)$	6
	a	1	1	2	2	$(2,2,1,1)$	$(1,1,2,2)$	6
	b	1	1	0	2	$(2,1,1,0)$	$(0,1,1,2)$	4
	c	2	1	2	0	$(2,2,1,0)$	$(0,1,2,2)$	5
$[\mu]$	ab	0	0	1	3	$(3,1,0,0)$	$(\mathbf{0},\mathbf{0},\mathbf{1},\mathbf{3})$	**4**
$[\mu]$	ac	1	0	3	1	$(3,1,1,0)$	$(0,1,1,3)$	5
$[\mu]$	bc	1	1	1	1	$(\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1})$	$(1,1,1,1)$	**4**
	abc	0	0	2	2	$(2,2,0,0)$	$(0,0,2,2)$	6

Table 3.1: Hamming distances from the formulas of the profile $\vec{\varphi}$ in Example 3.18 to each interpretation in the universe, together with the aggregated distances, for the leximax, leximin and sum aggregation functions. Models of the constraint μ are singled out: the optimal outcomes are the ones with overall minimal scores.

If d is a distance between interpretations and $\oplus$ is an aggregation function, the $(d, \oplus)$-*induced* $\mathcal{L}^n$*-assignment* $\preccurlyeq^{d,\oplus}$ *on interpretations* is obtained by taking $\preccurlyeq^{d,\oplus}(\vec{\varphi}) = \leq_{\vec{\varphi}}^{d,\oplus}$, for any $\mathcal{L}$-profile $\vec{\varphi}$. In the same vein, the $(d, \oplus)$-*induced* $\mathcal{L}^n$*-merging operator* $\Delta^{d,\oplus}$ is the operator induced by the $\mathcal{L}^n$-assignment $\preccurlyeq^{d,\oplus}$ on interpretations. This allows us to generate total, syntax insensitive m-faithful assignments.

Proposition 3.8 ([Konieczny and Pino Pérez, 2011])

If d is a distance between interpretations, $\oplus$ is an aggregation function and $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$ is an $\mathcal{L}$-profile, the $(d, \oplus)$-induced ranking $\leq_{\vec{\varphi}}^{d,\oplus}$ satisfies properties m_{1-9}, i.e., $\leq_{\vec{\varphi}}^{\oplus,\min}$ is a total preorder on interpretations that is syntax insensitive and is m-faithful.

Proposition 3.8 implies that the $(d, \oplus)$-induced $\mathcal{L}^n$-assignment $\preccurlyeq^{\oplus,\min}$ on interpretations is total, m-faithful and syntax insensitive, which, by Theorem 3.10, implies that the $(d, \oplus)$-induced merging operator $\Delta^{d,\oplus}$ satisfies postulates M_{1-6}.

Corollary 3.6

If d is a distance between interpretations and $\oplus$ is an aggregation function, the $(d, \oplus)$-induced revision operator $\Delta^{d,\oplus}$ satisfies postulates M_{0-8}.

Throughout this work we will typically focus on operators generated using Hamming distance d_{H} and drastic distance d_{D}, and the sum, leximax and leximin aggregation functions, denoted as $\Delta^{\mathrm{H},\oplus}$ and $\Delta^{\mathrm{D},\oplus}$, for $\oplus \in \{\mathrm{sum}, \mathrm{leximax}, \mathrm{leximin}\}$.

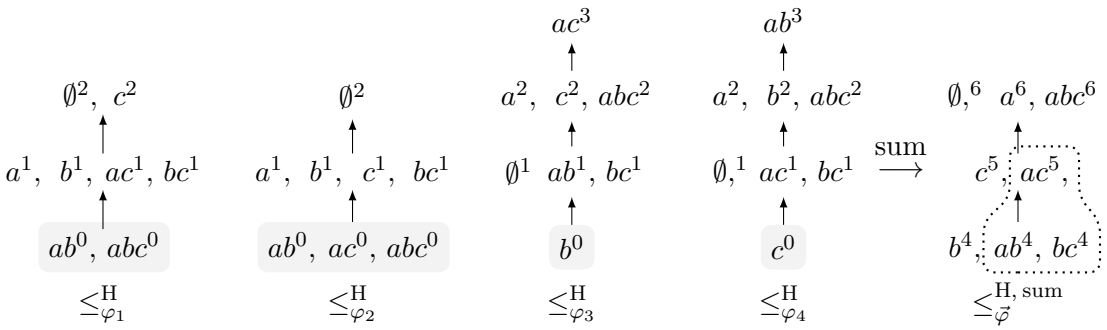


Figure 3.12: Total preorders $\leq^{\mathrm{H}}_{\varphi_i}$, for $i \in \{1,2,3,4\}$, corresponding to the opinions of the four Academy members in Example 3.18, as well as the aggregated preorder $\leq^{\mathrm{H,\,sum}}_{\vec{\varphi}}$, corresponding to the profile $\vec{\varphi}$. The models of the result are the most preferred models of the constraint μ (depicted here in the area bordered by the dotted line) in the collective preorder $\leq_{\vec{\varphi}}$. The superscripts next to each interpretation stand for distances; the sum aggregation function is written above the arrow separating the preorders in the profile from the collective preorder $\leq^{\mathrm{H,\,sum}}_{\vec{\varphi}}$.

Example 3.18: #OscarsSoFossilized, with distances

For the setting in Example 3.15, with $A = \{a, b, c\}$, the profile is $\vec{\varphi} = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$, with $\varphi_1 = a \wedge b$, $\varphi_2 = a \wedge (b \vee c)$, $\varphi_3 = \neg a \wedge b \wedge \neg c$. and $\varphi_4 = \neg a \wedge \neg b \wedge c$. The constraint is μ, with $\mu = (a \wedge b \wedge \neg c) \vee (a \wedge \neg b \wedge c) \vee (\neg a \wedge b \wedge c)$. The Hamming distances from formulas in the profile $\vec{\varphi}$ to each interpretation w, together with the aggregated distances according to the leximax, leximin and sum aggregation functions, are depicted in Table 3.1. The preorders $\leq^{\mathrm{H}}_{\varphi_i}$, as well as the aggregated preorder $\leq^{\mathrm{H,\,sum}}_{\vec{\varphi}}$, are depicted in Figure 3.12. We obtain that:

$$d_{\mathrm{H}}(\varphi_1, ab) = \min(d_{\mathrm{H}}(ab, ab), d_{\mathrm{H}}(abc, ab)) = \min(0, 1) = 0,$$

and that:

$$d^{\mathrm{sum}}_{\mathrm{H}}(\vec{\varphi}, ab) = d_{\mathrm{H}}(\varphi_1, ab) + d_{\mathrm{H}}(\varphi_2, ab) + d_{\mathrm{H}}(\varphi_3, ab) + d_{\mathrm{H}}(\varphi_4, ab) = 4.$$

With the other aggregation functions, we have that $d^{\mathrm{leximax}}_{\mathrm{H}}(\vec{\varphi}, ab) = (3, 1, 0, 0)$, i.e., the vector of distances from the formulas in $\vec{\varphi}$ to ab ordered in descending order, and $d^{\mathrm{leximin}}_{\mathrm{H}}(\vec{\varphi}, ab) = (0, 0, 1, 3)$, i.e., the vector of distances from the formulas in $\vec{\varphi}$ to ab ordered in ascending order. Note that $ab \approx^{\mathrm{H,\,sum}}_{\vec{\varphi}} bc$, since $d^{\mathrm{sum}}_{\mathrm{H}}(\vec{\varphi}, ab) = d^{\mathrm{sum}}_{\mathrm{H}}(\vec{\varphi}, bc)$. However, the situation is different when using the other aggregating functions: $bc <^{\mathrm{H,\,leximax}}_{\vec{\varphi}} ab$, since $(1, 1, 1, 1) <_{\mathrm{lex}} (3, 1, 0, 0)$, and $ab <^{\mathrm{H,\,leximin}}_{\vec{\varphi}} bc$, since $(0, 0, 1, 3) <_{\mathrm{lex}} (1, 1, 1, 1)$. We obtain that $[\Delta^{\mathrm{H,\,leximax}}_{\mu}(\vec{\varphi})] = \{bc\}$, $[\Delta^{\mathrm{H,\,leximin}}_{\mu}(\vec{\varphi})] = \{ab\}$ and $[\Delta^{\mathrm{H,\,sum}}_{\mu}(\vec{\varphi})] = \{ab, bc\}$.

The operators $\Delta^{\mathrm{H},\,\mathrm{sum}}$, $\Delta^{\mathrm{H},\,\mathrm{leximax}}$ and $\Delta^{\mathrm{H},\,\mathrm{leximin}}$ all embody different attitudes to the aggregation of information. Intuitively, $\Delta^{\mathrm{H},\,\mathrm{sum}}$ sees optimal outcomes in utilitarian terms and thereby favors the majority opinion, while $\Delta^{\mathrm{H},\,\mathrm{leximax}}$ attempts to improve the situation of the worse off agent, and usually veers towards egalitarian outcomes; the $\Delta^{\mathrm{H},\,\mathrm{leximin}}$ operator is elitist, in that it favors outcomes that improve the situation of the best off agent. On the other hand, the operators $\Delta^{\mathrm{D},\,\mathrm{sum}}$, $\Delta^{\mathrm{D},\,\mathrm{leximax}}$ and $\Delta^{\mathrm{D},\,\mathrm{leximin}}$, on the other hand, are all equivalent, i.e., $\Delta_{\mu}^{\mathrm{D},\,\mathrm{sum}}(\vec{\varphi}) \equiv \Delta_{\mu}^{\mathrm{D},\,\mathrm{leximin}}(\vec{\varphi}) \equiv \Delta_{\mu}^{\mathrm{D},\,\mathrm{leximax}}(\vec{\varphi})$, for any propositional profile $\vec{\varphi}$ and formula μ.

The difference between majoritarian and egalitarian operators can be hashed out in terms of the following postulates [Liberatore and Schaerf, 1998, Konieczny and Pino Pérez, 2011], to be thought of in conjunction with postulates M_{0-8} and applying for any $\mathcal{L}$-profiles $\vec{\varphi}_1$ and $\vec{\varphi}_2$, constraints μ, μ_1 and μ_2, and formulas φ_1 and φ_2:

($\mathrm{M}_{\mathsf{MAJ}}$) There exists an integer n such that $\Delta_{\mu}(\vec{\varphi}_1 + (\underbrace{\vec{\varphi}_2 + \cdots + \vec{\varphi}_2}_{n \text{ times}})) \models \Delta_{\mu}(\vec{\varphi}_2)$.

($\mathrm{M}_{\mathsf{ARB}}$) If $\Delta_{\mu_1}(\varphi_1) \equiv \Delta_{\mu_2}(\vec{\varphi}_2)$, $\Delta_{\mu_1 \leftrightarrow \neg\mu_2}(\varphi_1, \varphi_2) \equiv \mu_1 \leftrightarrow \neg\mu_2$, $\mu_1 \not\models \mu_2$ and $\mu_2 \not\models \mu_1$, then $\Delta_{\mu_1 \vee \mu_2}(\varphi_1, \varphi_2) \equiv \Delta_{\mu_1}(\varphi_1)$.

Postulate $\mathrm{M}_{\mathsf{MAJ}}$ says that a large enough coalition will sway the merging result in its favor, while postulate $\mathrm{M}_{\mathsf{ARB}}$ formalizes the idea that the median position is to be preferred [Liberatore and Schaerf, 1998]. We will not delve too much into these properties, except to say that operators $\Delta^{d,\,\mathrm{sum}}$ satisfy postulate $\mathrm{M}_{\mathsf{MAJ}}$ and operators $\Delta^{d,\,\mathrm{leximax}}$ satisfy postulate $\mathrm{M}_{\mathsf{ARB}}$.

3.5 Related work

Belief change in the sense relevant to us here begins, in earnest, with the AGM model of the 1980s [Alchourrón et al., 1985, Alchourrón and Makinson, 1985, Gärdenfors, 1988], with some of the main ideas going back, according to Peter Gärdenfors [Gärdenfors, 2011], slightly further [Harper, 1976, Levi, 1980].

The original AGM publications led to a watershed of works attempting to model belief change operators, and revision in particular, in more intuitive terms. Important proposals used entrenchment relations [Gärdenfors and Makinson, 1988], systems of spheres [Grove, 1988] and preorders on possible worlds [Katsuno and Mendelzon, 1992]. All of these models rely, in some way or another, on preferences, either among formulas or interpretations. The latter reference, of course, provides the basis for our own work, with a closely related model [Katsuno and Mendelzon, 1991] providing the basis for our presentation of update. It should be mentioned that the AGM model assumes a language that subsumes propositional logic, and, as such, is strictly more general than the Katsuno-Mendelzon model taken here as reference point. Nonetheless, the choice mechanisms that underly revision are similar across all representations: we chose the Katsuno-Mendelzon model

because, in our opinion, it exhibits these mechanisms in a way that is intuitive and that lends itself to applications across various other domains.

In the initial AGM model revision was often placed side by side with *contraction*, deemed equally important and analysed as a belief change operation in its own right. Contraction models removals of certain items of knowledge from a bigger corpus: in the terminology used here, contraction of φ with respect to μ would require changing φ in such a way that μ is not implied by the result. Though we have not looked at it here, contraction also affords an interpretation in terms of choice over interpretations [Caridroit et al., 2017]. In most formal models contraction and revision are intended to be inter-definable, with a sizeable literature devoted to finding solutions for when they are not [Delgrande, 2008, Delgrande and Wassermann, 2010, Delgrande and Wassermann, 2013, Zhuang et al., 2013, Zhuang et al., 2017].

Belief revision was understood early on to have many ideas in common with rational choice [Doyle, 1991, Rott, 1993, Schulte, 1999], with Hans Rott's book [Rott, 2001] providing an in-depth analysis of these connections, together with a set of bold philosophical claims. Most of the claims presented in Section 3.1 can be traced, in some way or another, to this work. Nonetheless, the formal model we work with is different, and the results, when not following directly from the Katsuno-Mendelzon paper [Katsuno and Mendelzon, 1992], were derived from scratch. There are more recent takes on the parallel between revision and rational choice [Bonanno, 2009, Arló-Costa and Pedersen, 2010, Hansson, 2014], but our impression is that the Katsuno-Mendelzon model remains the easiest one to work with. That postulates R_1, R_3 and R_{5-6} are essentially the same as the axioms for choice functions circulating in the literature on rational choice is, thus, news to no one, and it has even been argued that the equivalence is not a coincidence [Olsson, 2003].

Even Hans Rott's book, in all its comprehenesiveness, only looks at single-agent operations, which, in rational choice terms, is equivalent to individual decision makers. This is undoubtedly because at the time when Hans Rott was writing his book the framework for merging [Liberatore and Schaerf, 1998, Konieczny and Pino Pérez, 2002, Konieczny et al., 2004, Konieczny and Pino Pérez, 2011] had not yet been fully developed. As is clear by now, our view is that merging is to revision as social choice is to individual rational choice. There is nothing new about this either, with several publications attempting to build bridges between merging and social choice [Meyer, 2001, Meyer et al., 2001, Konieczny and Pino Pérez, 2005, Eckert and Pigozzi, 2005, Everaere et al., 2007, Everaere et al., 2014, Díaz and Pino Pérez, 2017, Everaere et al., 2017].

Our work on enforcement came out of an attempt to axiomatize enforcement in abstract argumentation [Baumann, 2012, Wallner et al., 2017], but the postulates took on a life of their own when formulated in propositional logic. The same postulates, we discovered later, had been used to model the dynamics of *desire* [Dubois et al., 2017], though the duality with revision embodied in Equations 3.1 and 3.2 had not been explicitly stated. Since enforcement does not guarantee full acceptance of the new information μ, it shares a similarity with *non-prioritized revision* [Hansson, 1999a, Hansson et al., 2001] and *belief promotion* operators [Schwind et al., 2018]. However, since the postulates characterizing

enforcement, and in particular, its success postulate E_2, are different, enforcement does not coincide with any of the proposals in this literature.

A useful comparison can be made with contraction, overlooked in this work but which is, as mentioned above, an important member of the belief change family. Contraction of a propositional formula φ by a propositional formula μ can be represented as a choice function using the identity $[\varphi] \cup \min_{\leq_\varphi}[\neg\mu]$, where $\leq_\varphi$ is a familiar, r-faithful preorder on interpretations that depends on φ [Caridroit et al., 2017]. In other words, the models of the contraction are obtained by adding the most plausible models of $\neg\mu$, according to $\leq_\varphi$, to the models of φ. Consider, now, the following example.

Example 3.19: Enforcement vs contraction

For a set of atoms $A = \{a, b\}$, take $[\varphi] = \{a\}$ and $[\mu] = \{b\}$. Then the result of enforcing μ with respect to φ is $[\varphi \triangleright \mu] = \varphi \vee \mu = \{a, b\}$, if the enforcement operator $\triangleright$ satisfies postulate E_2. On the other hand, contracting φ with respect to μ results in the set of interpretations:

$$\begin{aligned}[\varphi] \cup \min_{\leq_\varphi}[\neg\mu] &= \{a\} \cup \min_{\leq_\varphi}\{\emptyset, a, ab\} \\ &= \{a\} \cup \{a\} \\ &= \{a\}.\end{aligned}$$

The latter equality holds because $\leq_\varphi$ is assumed to satisfy the properties of an r-faithful assignment, such that $\min_{\leq_\varphi}\{\emptyset, a, ab\} = \{a\}$.

Thus, an agent who starts off believing that a is the true state of the world and queries a source that advocates for b will want to use an enforcement operator if the source is deemed credible enough. If, on the other hand, the source is deemed untrustowrthy and the agent wants to remove any information stemming from it, then it will use an contraction operator.

Note that the result of contraction in Example 3.19 is the same regardless of the particular preorder $\leq_\varphi$ used, as long as $\leq_\varphi$ satisfies the properties of an r-faithful assignment. Thus, Example 3.19 trades on what are uncontroversial cases for both enforcement and contraction, i.e., cases in which the result is unambiguously determined on the basis of the standard postulates alone. As such, it highlights the differences in how incoming information is treated by the two operations.

3.6 Conclusion

This chapter has introduced us to the main vehicles of belief change we will be studying throughout the rest of this work: revision, update, enforcement and merging. The defining characteristics of a belief change operator, we have seen, are the logical postulates used to axiomatize it, the preferences over outcomes that items of prior information are assumed to

influence, and the optimization behavior shown, through various representation theorems, to characterize belief change operators.

One of the aims of this chapter has been to show that belief change operations can be understood as choice procedures over the space of interpretations. The poster child for this approach is revision, which presents itself as a straightforward analogue to choice functions studied in rational choice theory [Sen, 1969, Sen, 1970, Grant and Zandt, 2009]. Theorem 3.1, in particular, tells us that an agent revising beliefs φ along the lines of postulates R_1 and R_{3-6} behaves as if it ranks outcomes in a total preorder $\leq_\varphi$, and always picks the minimal models of of the new information μ according to $\leq_\varphi$. Such an agent, then, behaves like a rational agent choosing the best elements from a given menu of options: the menu, here, consists of the models of μ, i.e., the possible worlds the agent is allowed to believe in light of new information, while the best elements are decided with reference to $\leq_\varphi$. Revision postulates R_1 and R_3 are equivalent to properties C_1 and C_2 of a choice function, as presented in Section 2.4, while postulates R_5 and R_6 are roughly equivalent to or properties C_3 and C_4, or properties α and β, as they are known in the theory of rational choice [Sen, 1969, Sen, 1970]. Postulate R_4, though it does not have an analogue in rational choice theory, reinforces the parallel by making sure that revision operators are not sensitive to the syntax of the formulas involved. Thus, taken together, postulates R_1 and R_{3-6} characterize choice functions over outcomes rationalizable by total preorders: accordingly, Theorem 3.1 aligns with standard choice theoretic results [Arrow, 1951, Sen, 1969]. The main difference between rational choice and revision, then, lies in the interpretation given to the concepts at play: a preference order, in rational choice, ranks items in terms of their desirability, whereas in belief change it ranks outcomes in terms of their plausibility.

In the wake of this result, we were able to make sense of other aspects of belief change through the lens of choice theory. Theorem 3.2 showed that the optimization behavior in Theorem 3.1 can be reproduced just as well with partial orders instead of total orders. Theorems 3.7 and 3.8 showed that update fits nicely into this perspective, with the choice being distributed across preorders induced by every model of the prior information φ rather than, as with revision, by φ as a whole. Section 3.3 introduced us to the novel type of belief change we called *enforcement*, with Theorem 3.9 showing that choice in the case of enforcement assumes a particular form: an enforcement operator has to figure out what models to add to the new information μ, rather than what models to discard. This is choice over outcomes that are not consistent with the new information, but choice nonetheless. Finally, Theorem 3.10 showed us that the choice perspective lends itself naturally to belief merging operators, as they can be seen as collective choice procedures.

In recasting belief change operators as choice procedures, postulate R_2, as well as its various incarnations, i.e., postulates U_2, E_2 or M_2, has been consistently put aside for separate treatment: this is because a belief change operator does not need it in order to function as a choice procedure. As Theorems 3.3 and 3.4 show, what postulate R_2, together with its avatars, does is to bias the choice relative to φ, by making sure that models of φ are given priority in the choice process. This is consistent with a view in

which outcomes consistent with a belief φ are considered the most plausible states of affairs, but raises the question as to what other attitudes towards these outcomes are reasonable. This is a question we will tackle in the next chapter.

CHAPTER 4

Revision as Biased Choice

Revision operators that satisfy postulate R_2 (besides the more standard postulates R_1 and R_{3-8}) can be understood to adopt a particular attitude towards prior information, which articulates the policy by which the agent's prior information behaves with respect to new data: if new information μ is consistent with existing beliefs φ, then the result of revision is simply $\varphi \wedge \mu$; in other words, the agent retains its initial beliefs and simply supplements them with the new item of information, if it can do so in a consistent way. Under the choice perspective of belief change we have been advocating, the agent ranks possible outcomes of the revision process in terms of their plausibility: in this setting, postulate R_2 makes sure that, when beliefs are up for grabs, models of the prior information φ are the first in line to be chosen. This attitude is in line with a view of revision according to which the prior information φ stands for the set of outcomes the agent finds most plausible, information not to be given up unless challenged by conflicting new data. This is a conservative attitude towards initial beliefs, guided by the desire to preserve them as much as possible. As pointed out in Section 3.1, it is lended support by Peter Gärdenfors' argument that information is not cheap and should be preserved to the best of one's ability [Gärdenfors, 1988, p.49], or by Harold Abelson's observation that humans treat beliefs as possessions [Abelson, 1986]. However, it is not the only attitude towards the prior information an agent can have.

Example 4.1: The art of diagnosis with biased preferences

A patient that has been previously diagnosed with asthma (a) sees two doctors, both of whom are aware of the patient's pre-existing condition. After a consultation it emerges that the patient is suffering from shortness of breath (b). The first doctor revises its beliefs by merely taking in this new information, i.e., concluding that the patient suffers from asthma and shortness of breath ($a \wedge b$). The second doctor infers that chest pain (c) must also be present ($a \wedge b \wedge c$): the two symptoms often go

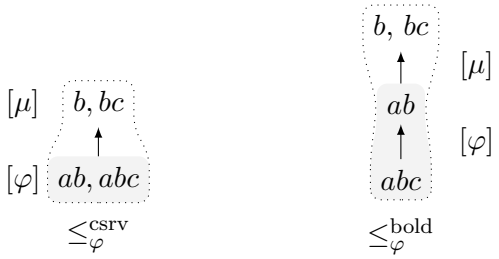


Figure 4.1: Two agents start with the same prior information $\varphi = a$ and receive the same new information $\mu = b$, but revise according to different preorders: $\leq_\varphi^{\text{csrv}}$ is guided by the more conservative imperative of preserving as much of the original information as possible; $\leq_\varphi^{\text{bold}}$ is bolder, in that it considers outcome abc more likely than ab, though both are consistent with φ, and draws the more specific, though riskier, conclusion. Models of φ are shaded in gray, models of μ are surrounded by the dotted line.

together in asthma, and the doctor is inclined to give added weight to this information. The conclusions of both doctors are based on accepting the new information, but they follow different strategies: the first doctor is more conservative in the way it uses the new information, whereas the second doctor draws a bolder conclusion.

Formalizing this as a belief change scenario, we would assume the alphabet $A = \{a, b, c\}$ and two agents, corresponding to the two doctors. Both agents are in possession of the same prior information $\varphi = a$ and they both revise by the same item of new information $\mu = b$. We assume that their revision policies are captured by two $\mathcal{L}$-revision operators $\circ^{\text{csrv}}$ and $\circ^{\text{bold}}$ that satisfy postulates R_1 and R_{3-6}, i.e., $\circ^{\text{csrv}}$ and $\circ^{\text{bold}}$ are exhaustive. We know, by Theorem 3.1, that this is equivalent to having the doctors rank outcomes according to a plausibility order specific to each, then settle on outcomes that are most plausible according to these plausibility orders. The first doctor concludes that the patient has asthma and shortness of breath, i.e., $\varphi \circ^{\text{csrv}} \mu = a \wedge b$. Note that $\varphi \circ^{\text{csrv}} \mu \equiv \varphi \wedge \mu$, i.e., the first doctor revises in accordance with postulate R_2. Since $[\varphi \wedge \mu] = \{ab, abc\}$, this revision policy is equivalent to saying that the agent considers outcomes ab and abc equally likely, and overall more likely than other models of μ, such as b or bc. Such an attitude is in accordance with an r-faithful $\mathcal{L}$-assignment on interpretations, and in particular with properties r_5 and r_7 in Section 3.1. A total preorder $\leq_\varphi^{\text{csrv}}$ consistent with this attitude is depicted on the left in Figure 4.1.

The second doctor concludes that the patient must have chest pain, alongside the already known asthma and shortness of breath, i.e., $\varphi \circ^{\text{bold}} \mu \equiv a \wedge b \wedge c$. Since $[\varphi \wedge \mu] = \{ab, abc\}$, this revision policy is equivalent to saying that the agent considers outcome abc more likely than every other model of μ, including ab. Such an attitude is not in accordance with an r-faithful $\mathcal{L}$-assignment on interpretations, and a total preorder $\leq_\varphi^{\text{bold}}$ consistent with this attitude is depicted on the right in Figure 4.1.

In particular, property r_5, saying that models of the prior information φ should be considered equally likely, is not satisfied by $\leq_\varphi^{\text{bold}}$. Nonetheless, in light of its experience, readings or hunches, the second doctor is happy to factor in information about the relative likelihood of certain outcomes, even if it means that they will be at variance with property r_5.

Example 4.1 shows two ways of approaching revision, based on two ways of ranking outcomes consistent with the prior information: a more conservative way, consistent with the familiar postulate R_2, and a bolder way, more eager to distinguish between such outcomes in terms of plausibility. The moral we want to draw from Example 4.1 is not that one of the strategies is better, or more rational, than the other, since we can, of course, come up with scnearios where either of the strategies will fare better than the other. We also want to resist the conclusion that the right way to model the difference between these cases is to show that one agent has access to more information than the other: in our setup both agents have access to the same primary information, φ and μ; the only thing that differs is the way in which they rank outcomes consistent with this information.

In this chapter we view the attitude embodied by the standard postulate R_2 as one among many that an agent can have towards its initial beliefs. By considering alternatives to postulate R_2, we are able to axiomatize revision operators that embody a wider range of attitudes towards prior information, and characterize these operators in terms of the types of preorders they induce on the set of possible worlds. To illustrate these principles we provide concrete operators, constructed using the ingredients introduced in Section 2.3: a notion of *distance* between interpretations and an *aggregation function* that ranks possible worlds depending on the initial beliefs. We also show, in each case, how these operators fit into the landscape of new postulates introduced. Without the theoretical apparatus of the new postulates, the concrete operators put forward would be merely classified as deviant, since they do not satisfy the traditional postulate R_2. But through the present analysis they can be viewed as encoding distinct and characterizable stances an agent can take towards its beliefs.

4.1 Postulates for biased revision operators

In Section 3.1 we presented a set of postulates for revision, which we divided into several groups. Postulates R_1 and R_{3-6} defined *exhaustive* revision operators, while postulates R_1, R_{3-5} and R_{7-8} defined *exclusive* revision operators. In this chapter we will focus on exhaustive operators, which, according to Theorem 3.1, are represented by total, syntax insensitive $\mathcal{L}$-assignments on interpretations. The aim here is to hold postulates R_1 and R_{3-6} fixed and explore alternatives to postulate R_2. We will do this by considering weaker versions of postulate R_2, as well as postulates that express related, but ultimately different intuitions.

Thus, we put forward the following postulates, meant to apply to any propositional formulas φ and μ, and complete propositional formulas $\dot{\varphi}$:

(R_9) If $\varphi \wedge \mu$ is consistent, then $\varphi \circ \mu \models \varphi \wedge \mu$.

(R_{10}) If $\varphi \wedge \mu$ is consistent, then $\varphi \wedge \mu \models \varphi \circ \mu$.

(R_{11}) If $\varphi \circ \mu \models \overline{\varphi}$, then $\varphi \circ \mu \equiv \mu$.

(R_{12}) If $\mu \not\models \overline{\varphi}$, then $(\varphi \circ \mu) \wedge \overline{\varphi}$ is inconsistent.

(R_{13}) If $\dot{\varphi} \models \varphi \circ \mu$, then $\dot{\varphi} \models (\varphi \vee \dot{\varphi}) \circ \mu$.

(R_{NEUT}) $\rho(\varphi \circ \mu) \equiv \rho(\varphi) \circ \rho(\mu)$.

Each of these postulates encodes a particular type of attitude towards prior information, and they are intended to be thought of in conjunction with the basic set of postulates R_1 and R_{3-6}. Postulate R_9 models an agent who reserves the right to drop information from φ if it sees fit to, even if that information is consistent with μ: we may imagine this is done on the basis of certain preferences over the information encoded by φ, i.e., the agent is partial towards some of the models of φ to the detriment of others, along the lines of the second doctor in Example 4.1. This type of discrimination can be explained by the agent having some background knowledge of the relative likelihoods of certain outcomes, as was the case in Example 4.1, or be the result of some heuristic that the agent uses to process information.

Example 4.2: Steve

Consider Steve, the subject of a classical example by Daniel Kahneman and Amos Tversky:

> An individual has been described by a neighbor as follows: "Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail." Is Steve more likely to be a librarian or a farmer?" [Kahneman, 2011]

Most people, we are told, reply that Steve is more likely to be a librarian: they do so based primarily on the stereotypical image of what it is to be a librarian, while disregarding more useful facts, such as the statistical distribution of librarians versus farmers in the population, which would favor farmers. Humans, it seems, readily use a representativeness heuristic to draw conclusions that are otherwise unwarranted.

Imagine this example simplified to fit the parameters of a quick revision scenario: an agent knows that Steve is shy (a) and learns, in addition, that Steve is helpful and very well organized (b). The agent concludes that Steve is a librarian (c). Formally,

this example has the same structure as Example 4.1, with prior information $\varphi = a$, new information $\mu = b$ and conclusion $\varphi \circ \mu \equiv a \wedge b \wedge c$, and results in the same observation: the agent considers the outcome abc more likely than outcome ab. But this time it is on the basis of an availability bias: outcome abc is simply more salient in the agent's mind than ab alone, even though they are both consistent with φ and μ. Who are we to judge?

Postulate R_{10} refers to an attitude that is, in some ways, the oposite of the attitude in postulate R_9: it models an agent who incorporates all information in $\varphi \wedge \mu$, and possibly extends this to cover more ground. Taken together, postulates R_9 and R_{10} imply that $\varphi \circ \mu$ is equivalent to $\varphi \wedge \mu$, when $\varphi \wedge \mu$ is consistent, i.e., they are equivalent to the classical postulate R_2.

Postulates R_{11} and R_{12} focus on the dual formula $\overline{\varphi}$ obtained by replacing every literal in φ with its dual, i.e., negated version (see Section 2.1). Why are these formulas significant? Note that in certain special cases, such as when φ is a conjunction of literals or a complete formula (i.e., with exactly one model), then $\overline{\varphi}$ will be a formula whose models are complements of the models of φ.

Example 4.3: Dual formulas as opposite points of view

For the set of atoms $A = \{a, b, c\}$, consider formulas $\varphi_1 = a \wedge b \wedge \neg c$, $\varphi_2 = a \wedge b$, $\varphi_3 = a \vee b$, $\varphi_4 = b \rightarrow c$. We have that $\overline{\varphi_1} = \neg a \wedge \neg b \wedge c$ and $\overline{\varphi_2} = \neg a \wedge \neg b$, with $[\varphi_1] = \{ab\}$ and $[\varphi_2] = \{ab, abc\}$ while $[\overline{\varphi_1}] = \{c\}$ and $[\overline{\varphi_2}] = \{\emptyset, c\}$. On the other hand, $[\varphi_3] = \{a, b, ab, ac, bc, abc\}$ and $[\varphi_4] = \{\emptyset, a, c, ac, bc, abc\}$, while $[\overline{\varphi_3}] = \{\emptyset, a, b, c, ac, bc\}$ and $[\overline{\varphi_4}] = \{\emptyset, a, b, ab, bc, abc\}$. Note that $[\varphi_1] \cap [\overline{\varphi_1}] = \emptyset$ and $[\varphi_2] \cap [\overline{\varphi_2}] = \emptyset$, but $[\varphi_3] \cap [\overline{\varphi_3}] = \{a, b, ac, bc\}$ and $[\varphi_4] \cap [\overline{\varphi_4}] = \{\emptyset, a, bc, abc\}$.

Conjunctions of literals or complete formulas can be thought of as being very specific, in the sense that they model agents with definite opinions on some (or all) atoms. In this case, $\overline{\varphi}$ can be thought of as the point of view opposite to that of φ, i.e., the outcome least likely to be true if φ is. As Example 4.3 illustrates, however, this analogy breaks down if φ is a different type of formula, such as a disjunction or a conditional. In this case φ and $\overline{\varphi}$ can share models, and the distinction between a formula and its dual is not as clear-cut anymore. The claim, then, is only that there are situations in which it makes sense to view φ and $\overline{\varphi}$ as embodying opposing stances, and situations can be imagined in which it is desirable to put bounds on the revision function in terms of how it treats information encoded by $\overline{\varphi}$. This is the case if the agent has, or is required to have, a definite opinion on every item from an agenda, as is typically the case in Judgment Aggregation [Endriss, 2016]; if φ is a 'vivid' formula [Levesque, 1986]; or, if it encodes something like an agent's preferred bundle from a set of available items. In all these cases φ can be required to be a conjunction of literals or a complete formula. In this context, postulate R_{11} says that if φ undergoes revision by a formula μ embodying

such an adverse perspective, then the agent must adopt μ: in other words, the agent has no room for maneuvering towards a more amenable middle ground. Such a revision policy makes more sense when considered alongside postulate R_{12}, which specifies that if the agent has the option of believing states of affairs *not* compatible with $\overline{\varphi}$, it should wholeheartedly adopt those as the most plausible stance. Taken together, postulates R_{11} and R_{12} inform the agent to believe states of affairs compatible with $\overline{\varphi}$ only if it has no other choice in the matter: the models of $\overline{\varphi}$ should be part of a viewpoint one is willing to accept only as a last resort.

Example 4.4: Recommending as revising

Consider an online streaming platform that gathers data about its users in order to tailor recommendations to their likes and preferences. Suppose φ encodes something like the information this system has about a specific user (e.g., the items that the user liked), μ represents a query from the user and $\varphi \circ \mu$ represents the set of results suggested to the user by the online platform, in response to the query μ.

We can imagine that the platform uses its knowledge φ to construct a profile of the user, which then serves as guide about what to recommend: in revision terms, this profile serves as the revision policy, or, as we will soon see, a ranking of the possible outcomes. We can also imagine that it is important for the platform that this profile contains, besides the items that the user is most likely to appreciate, also a list of items that the user will *dislike*, so as to avoid suggesting those as much as possible (i.e., unless the user explicitly requests them). In revision terms, we can conceptualize this as a set of interpretations that are in the result $\varphi \circ \mu$ only as a last resort, and it is not difficult to see that the duals of the 'liked' outcomes are good candidates for these, likely to be loathed, options.

Postulate R_{13} enforces coherence of change when the prior information includes some element that would have been chosen in a different state of mind, and is best understood through an example.

Example 4.5: So many options

An agent intends to go to an art museum (a), the beach (b) and a concert (c), encoded as the complete propositional formula $\varphi = a \wedge b \wedge c$, with the set of atoms $A = \{a, b, c\}$. The agent then learns that it only has time for one of these activities: this is encoded as newly acquired information $\mu = (a \wedge \neg b \wedge \neg c) \vee (\neg a \wedge b \wedge \neg c) \vee (\neg a \wedge \neg b \wedge c)$, with $[\mu] = \{a, b, c\}$. The agent chooses the art museum a, i.e., $\varphi \circ \mu \equiv \dot{\varphi} \equiv a \wedge \neg b \wedge \neg c$. Note that $\dot{\varphi} \models \varphi \circ \mu$. If the agent's initial intentions had been more inclusive in the sense that it would have expressed a willingness to do either all three activities or just visit the museum, encoded as $\varphi \vee \dot{\varphi}$, then, faced with the same new information μ, the option of going to the art museum (i.e., $a \wedge \neg b \wedge \neg c$) should still feature as one of its most preferred options.

We have also found it suitable to add here a neutrality postulate $\mathrm{R}_{\mathsf{NEUT}}$, requiring that a revision operator does not favor propositional atoms based solely on their names. This idea is expressed by requiring the revision output to be invariant under a renaming ρ of atoms, and, in conjunction with the insensitivity to syntax postulate R_4, it is perhaps natural to expect it from any revision operator. The inspiration for postulate $\mathrm{R}_{\mathsf{NEUT}}$ is the social choice literature and, though it has appeared in belief change before, under various guises [Herzig and Rifi, 1999, Marquis and Schwind, 2014, Haret et al., 2016b], neutrality usually goes unstated in standard presentations of revision.

4.2 Biased preferences over outcomes

A clearer view of postulates R_{9-13} and $\mathrm{R}_{\mathsf{NEUT}}$ emerges when looking at the constraints they impose on how models of φ are placed in the preorder $\leq_\varphi$. That is to say, we assume an agent with prior information φ ranks interpretations in a total preorder $\leq_\varphi$, i.e., that there exists a total, syntax insensitive $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that satisfies properties r_{1-4}, as presented in Section 3.1. The additional constraints we want to consider in this chapter are expressed in the following properties, applying for any interpretations w_1, w_2 and v and propositional formulas φ and ε_v, where $[\varepsilon_v] = \{v\}$:

(r_7) If $w_1 \in [\varphi]$ and $w_2 \notin [\varphi]$, then $w_1 <_\varphi w_2$.

(r_8) If $w_1 \in [\varphi]$, then $w_1 \leq_\varphi w_2$.

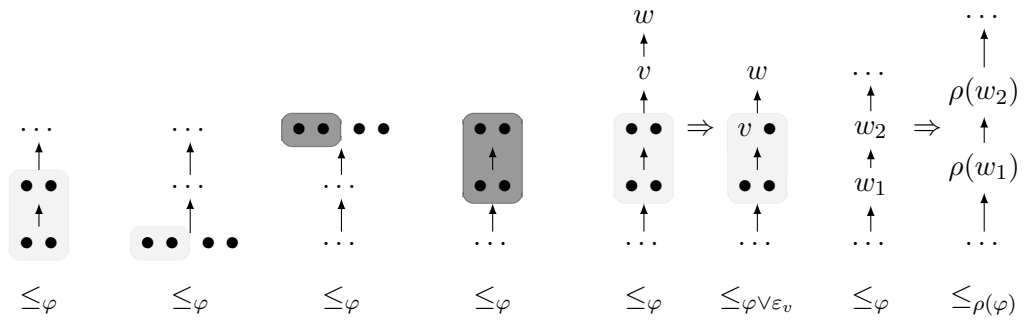
(r_9) If $w_1 \in [\overline{\varphi}]$, then $w_2 \leq_\varphi w_1$.

(r_{10}) If $w_1 \in [\overline{\varphi}]$ and $w_2 \notin [\overline{\varphi}]$, then $w_2 <_\varphi w_1$.

(r_{11}) If $v \leq_\varphi w$, then $v \leq_{\varphi \vee \varepsilon_v} w$.

($\mathrm{r}_{\mathsf{NEUT}}$) If $w_1 \leq_\varphi w_2$, then $\rho(w_1) \leq_{\rho(\varphi)} \rho(w_2)$.

Each of properties r_{7-11} tells us something about how prior information φ biases a plausibility ordering over outcomes, especially with regards to the outcomes consistent with φ (i.e., models of φ), whereas the neutrality property $\mathrm{r}_{\mathsf{NEUT}}$ tells us something about the kind of bias the ordering should avoid. A schematic view of these preorders is offered in Figure 4.2. Property r_7, depicted in Figure 4.2-(a), says that models of φ are generally considered more plausible than interpretations that do not satisfy φ, but models of φ themselves may not be equally plausible relative to each other. Property r_7 is familiar from Section 3.1, though there it was always considered in conjunction with r_5 or r_6, and never by itself. Here we shine a spotligh on it alone. Property r_8, depicted in Figure 4.2-(b), says that models of φ are minimal elements in φ, though possibly not uniquely so. Note that property r_8 implies property r_6 (see Section 3.1), which says that models of φ should be equally plausible in $\leq_\varphi$: however, r_8 tells us more than r_5: it tells us that the agent not only considers models of φ as equally plausible, but also at least



(a) r_7 (b) r_8 (c) r_9 (d) r_{10} (e) r_{11} (e) r_{NEUT}

Figure 4.2: Schematic view of prototypical preorders satisfying each of the properties r_{6-10}; models of φ are in the light gray area, models of $\overline{\varphi}$ are in the dark gray area.

as plausible as any other outcome. Properties r_{9-10}, depicted in Figure 4.2-(c,d), say that models of the dual formula φ are the least plausible interpretations in $\leq_\varphi$, while property r_{11}, depicted in Figure 4.2-(e), says that if v is more plausible than w when the initial beliefs are φ, then v would still be more plausible than w if it were part of the initial beliefs. Lastly, the neutrality property r_{NEUT} says that the ranking $\leq_\varphi$ should be invariant when renaming the atoms.

Note that properties r_{7-8}, together, are equivalent to property r_6, presented in Section 3.1. Thus, properties r_{7-8} together with the standard properties r_{1-5} define what is more commonly known as a *faithful assignment* [Katsuno and Mendelzon, 1992], or, as we have named it here, a total, syntax independent r-faithful $\mathcal{L}$-assignment on interpretations. Such an assignment places all and only models of φ on the lowest level of $\leq_\varphi$, and corresponds to an agent for which outcomes consistent with its prior belief are the most plausible states of affairs. This attitude, as is apparent here, arises out of a combination of two attitudes that can be looked at separately.

Properties r_{7-11} and r_{NEUT} turn out to characterize postulates R_{9-13} and R_{NEUT} on the semantic level, as per the following representation result. Recall that an $\mathcal{L}$-revision operator $\circ$ is represented by an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations if $[\varphi \circ \mu] = \min_{\leq_\varphi}[\mu]$, for any propositional formulas φ and μ, and that, by Theorem 3.1, if $\circ$ is exhaustive, then there always exists a total assignment $\preccurlyeq$ (i.e., such that $\leq_\varphi$ is a total preorder) representing it. Recall, as well, that the $\mathcal{L}$-proxy of a pair $\{w_1, w_2\}$ of interpretations is a propositional formula $\varepsilon_{1,2}$ for which $[\varepsilon_{1,2}] = \{w_1, w_2\}$.

Theorem 4.1

If $\circ$ is revision operator that satisfies postulates R_1 and R_{3-6} (i.e., is exhaustive) and $\preccurlyeq$ is a total, syntax insensitive $\mathcal{L}$-assignment on interpretations that represents it, then the following equivalences hold:

(1) $\circ$ satisfies postulate R_9 iff $\preccurlyeq$ satisfies property r_7;

(2) $\circ$ satisfies postulate R_{10} iff $\preccurlyeq$ satisfies property r_8;

(3) $\circ$ satisfies postulate R_{11} iff $\preccurlyeq$ satisfies property r_9;

(4) $\circ$ satisfies postulate R_{12} iff $\preccurlyeq$ satisfies property r_{10};

(5) $\circ$ satisfies postulate R_{13} iff $\preccurlyeq$ satisfies property r_{11};

(6) $\circ$ satisfies postulate R_{NEUT} iff $\preccurlyeq$ satisfies property r_{NEUT}.

Proof

We start with Equivalence 1 and show each direction in turn.

("$\Leftarrow$") Take a total $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations satisfying property r_7 and the revision operator $\circ$ represented by it. We want to show that $\circ$ satisfies postulate R_9, and start by assuming that $\varphi \wedge \mu$ is consistent. This implies that μ is consistent and, by postulate R_3, that $\varphi \circ \mu$ is consistent as well. Take, then, an interpretation w such that $w \in [\varphi \circ \mu]$, and suppose $w \notin [\varphi \wedge \mu]$. Since $[\varphi \circ \mu] = \min_{\leq_\varphi}[\mu]$, we obtain that $w \in \min_{\leq_\varphi}[\mu]$ and hence $w \in [\mu]$. Thus, the fact that $w \notin [\varphi \wedge \mu]$ implies that $w \notin [\varphi]$. But, by assumption, it holds that $[\varphi \wedge \mu] \neq \emptyset$, which means that there exists $w' \in [\varphi \wedge \mu]$ and, by property r_7, it follows that $w' <_\varphi w$. But we also have that $w \in \min_{\leq_\varphi}[\mu]$: since $\leq_\varphi$ is total, this implies that $w \leq_\varphi w'$: we have arrived at a contradiction.

("$\Rightarrow$") Take an exhaustive revision operator $\circ$ additionally satisfying postulate R_9 and a total $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that represents it. To show that $\leq_\varphi$ satisfies property r_7, take interpretations w_1 and w_2 such that $w_1 \in [\varphi]$ and $w_2 \notin [\varphi]$. We then have that $\varphi \wedge \varepsilon_{1,2}$ is consistent and hence, by postulate R_9, that $\varphi \circ \varepsilon_{1,2} \models \varphi \wedge \varepsilon_{1,2}$. Since $[\varphi \circ \varepsilon_{1,2}]$ is, by postulates R_1 and R_3, a non-empty subset of $[\varepsilon_{1,2}] = \{w_1, w_2\}$, we have that at least one of w_1 and w_2 is in $[\varphi \circ \varepsilon_{1,2}]$. Notice, now, that we cannot have $w_2 \in [\varphi \circ \varepsilon_{1,2}]$, since it would follow that $w_2 \in [\varphi \wedge \varepsilon_{1,2}]$ and, *a fortiori*, that $w_2 \in [\varphi]$, which contradicts a previous finding. Thus, $[\varphi \circ \varepsilon_{1,2}] = \{w_1\}$. Since $\leq_\varphi$ represents $\circ$, we have that $[\varphi \circ \varepsilon_{1,2}] = \min_{\leq_\varphi}[\varepsilon_{1,2}] = \{w_1\}$, i.e., that $w_1 <_\varphi w_2$.

For Equivalence 2, we show again each direction in turn.

("$\Leftarrow$") Take, first, a total $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations satisfying property r_8, and the revision operator $\circ$ represented by it. Assuming that $\varphi \wedge \mu$ is consistent, we want to show that for any $w_1 \in [\varphi \wedge \mu]$, it holds that $w_1 \in [\varphi \circ \mu]$ as well. Since $\circ$ is represented by $\preccurlyeq$, this is equivalent to showing that $w \in \min_{\leq_\varphi}[\mu]$. Take, then, an interpretation $w \in [\varphi \wedge \mu]$, and an arbitrary interpretation $w' \in [\mu]$. Applying property r_8, we infer that $w_1 \leq_\varphi w_2$, which then implies that $w \in \min_{\leq_\varphi}[\mu]$.

("⇒") Take an exhaustive revision operator $\circ$ that additionally satisfies postulate R_{10}, and a total $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that represents it. To show that $\leq_\varphi$ satisfies property r_8, take two interpretations w_1 and w_2 such that $w_1 \in [\varphi]$. Then, by postulate R_{10}, we have that $\varphi \wedge \varepsilon_{1,2} \models \varphi \circ \varepsilon_{1,2}$. This implies that $w_1 \in [\varphi \circ \varepsilon_{1,2}]$, i.e., that $w_1 \in \min_{\leq_\varphi}[\varepsilon_{1,2}]$. Thus, it holds that $w_1 \leq_\varphi w_2$.

Equivalences 3 and 4 are analogous to 1 and 2, respectively. For Equivalence 5, assume first that postulate R_{13} holds, and take interpretations w and v and a formula $\dot{\varphi}$ such that $v \leq_\varphi w$ and $[\dot{\varphi}] = \{v\}$. To show that $v \leq_{\varphi \vee \dot{\varphi}} w$, we must show that $v \in [(\varphi \vee \dot{\varphi}) \circ \varepsilon_{v,w}]$, where $\varepsilon_{v,w}$ is a formula such that $[\varepsilon_{v,w}] = \{v, w\}$. This follows immediately by applying postulate R_{13}. Conversely, suppose $[\dot{\varphi}] = \{v\}$, and take $w \in [\varphi \circ \mu]$. Then, we get that $v \leq_\varphi w$, and we can apply property r_{11} to derive the conclusion.

For Equivalence 6, recall that applying a renaming ρ to a set $\mathcal{W}$ of interpretations simply applies ρ to every interpretation in $\mathcal{W}$, such that if $w \in \mathcal{W}$, then it holds that $\rho(w) \in \rho(\mathcal{W})$. We will also make extensive use of Proposition 2.2 from Section 2.1, saying that the renaming functions commutes across the semantic line, i.e., $[\rho(\varphi)] = \rho([\varphi])$, for any propositional formula φ. We again take each direction in turn.

("⇒") Take an exhaustive revision operator $\circ$ that additionally satisfies postulate R_{NEUT}, and a total $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that represents it. Take interpretations w_1 and w_2 and suppose that $w_1 \leq_\varphi w_2$. Then $w_1 \in [\varphi \circ \varepsilon_{1,2}]$, and hence $\rho(w_1) \in \rho([\varphi \circ \varepsilon_{1,2}])$. By Proposition 2.2, we obtain that $\rho(w_1) \in [\rho(\varphi \circ \varepsilon_{1,2})]$ and by postulate R_{NEUT} it follows that $\rho(w_1) \in [\rho(\varphi) \circ \rho(\varepsilon_{1,2})]$. This implies that $\rho(w_1) \in \min_{\leq_{\rho(\varphi)}}[\rho(\varepsilon_{1,2})]$, which by Proposition 2.2 implies that $\rho(w_1) \in \min_{\leq_{\rho(\varphi)}}\rho([\varepsilon_{1,2}])$. Thus, $\rho(w_1) \leq_{\rho(\varphi)} \rho(w_2)$.

("⇐") Take a total $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations that satisfies property r_{NEUT} and the revision operator $\circ$ represented by it. By Proposition 2.2 we have that $[\rho(\varphi \circ \mu)] = \rho([\varphi \circ \mu]) = \rho(\min_{\leq_\varphi}[\mu])$, and $[\rho(\varphi) \circ \rho(\mu)] = \min_{\leq_{\rho(\varphi)}}[\rho(\mu)] = \min_{\leq_{\rho(\varphi)}}\rho([\mu])$. We show that $[\rho(\varphi \circ \mu)] = [\rho(\varphi) \circ \rho(\mu)]$ by double inclusion. Take, first, $\rho(w_1) \in \rho(\min_{\leq_\varphi}[\mu])$, for $w_1 \in \min_{\leq_\varphi}[\mu]$, and $\rho(w_2) \in \rho([\mu])$, for $w_2 \in [\mu]$. Then $w_1 \leq_\varphi w_2$ and by r_{NEUT} we get that $\rho(w_1) \leq_{\rho(\varphi)} \rho(w_2)$, which implies that $\rho(w_1) \in \min_{\leq_{\rho(\varphi)}}\rho([\mu])$. This shows that $[\rho(\varphi \circ \mu)] \subseteq [\rho(\varphi) \circ \rho(\mu)]$. Next, take $\rho(w_1) \in \min_{\leq_{\rho(\varphi)}}\rho([\mu])$, for $w_1 \in [\mu]$, and $\rho(w_2) \in \rho([\mu])$, for $w_2 \in [\mu]$. We get that $\rho(w_1) \leq_{\rho(\varphi)} \rho(w_2)$, which, via property r_{NEUT} and the renaming ρ^{-1}, implies that $w_1 \leq_\varphi w_2$. Thus, $w_1 \in \min_{\leq_\varphi}[\mu]$ and hence $\rho(w_1) \in \rho(\min_{\leq_\varphi}[\mu])$.

Note that properties r_7 and r_8, together, imply that the models of φ in a total preorder $\leq_\varphi$ are on the bottom. If this happens for every propositional formula φ, this implies that the overall assignment is r-faithful. In other words, Equivalences 1 and 2 from Theorem 4.1, added to Theorem 3.1, make up the classical representation result for

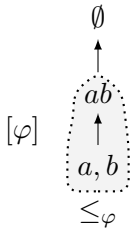


Figure 4.3: An agent with prior beliefs φ, $[\varphi] = \{a, b, ab\}$, thinks outcomes a and b are more likely than ab. When revising by φ, the result is $[\varphi \circ \varphi] = \{a, b\}$, which does not fit with postulate $\mathrm{R}_{\mathsf{IAHB}}$.

revision operators, which appears here as Theorem 3.5. Here we have opted for a more fine-grained approach to the placement of models of φ in $\leq_\varphi$, which allows a more diverse representation of the different types of attitudes an agent can have towards initial beliefs.

4.3 Indifference to already held beliefs

One particular consequence of weakening postulate R_2 is that the following property, called here $\mathrm{R}_{\mathsf{IAHB}}$, for *indifference to already held beliefs*, is not guaranteed to hold anymore. We write this property as a standalone postulate, meant to apply for any propositional formula φ:

($\mathrm{R}_{\mathsf{IAHB}}$) $\varphi \circ \varphi \equiv \varphi$.

Postulate $\mathrm{R}_{\mathsf{IAHB}}$ says that revising with information the agent already believes does not change the agent's prior beliefs, and is an instance of a more general principle, which is already implied by the standard postulates R_{1-6}, ensuring that revision by any formula μ such that $\varphi \models \mu$ results in φ.

Quick reflection reveals that postulate R_2 implies $\mathrm{R}_{\mathsf{IAHB}}$, though the converse does not hold, and neither of the weaker postulates R_9 or R_{10}, individually, manages to guarantee $\mathrm{R}_{\mathsf{IAHB}}$. The plausibility ranking view of revision proves useful in understanding postulate $\mathrm{R}_{\mathsf{IAHB}}$: even if the agent ranks models of φ as overall more plausible than other interpretations, if they are allowed to rank models of φ unequally, then $\mathrm{R}_{\mathsf{IAHB}}$ is not guaranteed to hold.

Example 4.6: Independence to already held beliefs

For the set of atoms $A = \{a, b\}$, consider an agent whose prior beliefs are represented by the formula $\varphi = a \vee b$, who revises according to a revision operator $\circ$ that satisfies postulate R_9, and who ranks interpretations according to the total preorder $\leq_\varphi$ in Figure 4.3. The agent finds out, on good authority, information consistent with φ.

On revising by φ, the result is $[\varphi \circ \varphi] = \min_{\leq_\varphi}[\mu] = \{a, b\}$, i.e., $\varphi \circ \varphi \equiv (a \leftrightarrow \neg b)$, and it is apparent that $\varphi \circ \varphi \not\equiv \varphi$ and that postulate $\mathrm{R}_{\mathsf{IAHB}}$ is not satisfied.

In fact, postulate $\mathrm{R}_{\mathsf{IAHB}}$ characterizes a property on interpretations that coincides with neither of the properties introduced in this chapter, but is familiar from Section 3.1: recall, from there, property r_5, saying that if two interpretations w_1 and w_2 are models of φ, then $w_1 \approx_\varphi w_2$.

Theorem 4.2

If an $\mathcal{L}$-revision operator $\circ$ satisfies postulates R_1 and R_{3-6} (i.e., is exhaustive) and $\preccurlyeq$ is a total $\mathcal{L}$-assignment on interpretations that represents it, then $\circ$ satisfies postulate $\mathrm{R}_{\mathsf{IAHB}}$ if and only if $\preccurlyeq$ satisfies property r_5.

Proof

("$\Rightarrow$") Consider an exhaustive revision operator $\circ$ that satisfies postulates R_1 and R_{3-6} and, in addition, postulate $\mathrm{R}_{\mathsf{IAHB}}$, and an assignment $\preccurlyeq$ that represents it. Take two interpretations v_1 and v_2 such that $v_1, v_2 \in [\varphi]$. Applying postulate $\mathrm{R}_{\mathsf{IAHB}}$, it follows that $v_1, v_2 \in [\varphi \circ \varphi]$, and hence $v_1, v_2 \in \min_{\leq_\varphi}[\varphi]$, i.e., $v_1 \approx_\varphi v_2$.

("$\Leftarrow$") Starting from an assignment $\preccurlyeq$ and the revision operator that represents it, assuming that r_5 is satisfied, we have that $\min_{\leq_\varphi}[\varphi] = [\varphi]$, which implies that postulates $\mathrm{R}_{\mathsf{IAHB}}$ is satisfied.

The ability to distinguish among models of one's prior beliefs in terms of plausibility points to a more graded view of what it means to believe φ. Thus, an agent might have a certain threshold of plausibility, along the lines of what is known in epistemology as *the Lockean thesis* [Foley, 1993], according to which it calibrates its beliefs: anything above the threshold counts as part of the belief φ and anything below counts as disbelief. This fits with the idea that an agent might assign different degrees of plausibility to states of affairs consistent with its belief φ: indeed, this is the point of view we endorse here, in contrast to more standard approaches, which consider that an agent assigns equal degrees of plausibility to all items of its belief. Thus, incoming information that confirms an agent's belief might have the effect of *reinforcing* parts that are given more plausibility at the expense of parts that are given less, and this is the kind of situation we take to be modeled by Example 4.6.

What would be worrying would be a revision policy that makes an agent cycle between different viewpoints when confronted repeatedly with the same type of information: we will see that for revision operators satisfying R_8 this concern is unwarranted, but we must first introduce some new notation. If φ is a propositional formula and $\circ$ is a revision operator, then φ^i is the formula obtained by revising φ by itself, using $\circ$, an i number of

times. Thus, $\varphi^0 = \varphi$ and $\varphi^{i+1} = \varphi^i \circ \varphi$. Consider now the following property, written as a postulate meant to apply for any propositional formula φ:

($\mathrm{R}_{\mathsf{STAB}}$) There is $n \geq 1$ such that $\varphi^m \equiv \varphi^n$, for every $m \geq n$.

Postulate $\mathrm{R}_{\mathsf{STAB}}$, where '$\mathsf{STAB}$' stands for *Stability*, implies that repeated revision by φ ultimately settles (or stabilizes) on a set of models that does not change through subsequent revisions by φ. A revision operator $\circ$ is *stable* if it satisfies postulate $\mathrm{R}_{\mathsf{STAB}}$. The following result proves relevant to the issue of stability.

Proposition 4.1

If a revision operator $\circ$ satisfies postulates R_1 and R_9, then $\varphi^{i+1} \models \varphi^i$.

Proof

By postulate R_1, we have that $\varphi \circ \varphi \models \varphi$, and thus $\varphi^1 \models \varphi^0$. Applying postulate R_9, we have that $(\varphi \circ \varphi) \circ \varphi \models (\varphi \circ \varphi) \wedge \varphi \models \varphi \circ \varphi$. Thus, $\varphi^2 \models \varphi^1$, and it is straightforward to see how this argument is iterated to get the conclusion.

If the operator $\circ$ also satisfies postulate R_3 (i.e., if the revision formula is consistent, then the revision result is also consistent), it follows that if φ is consistent, then φ_i is consistent, for any $i \geq 0$. Thus, combining this fact and Proposition 4.1, we get that repeated revision by φ leads to a chain of ever more specific formulas, i.e., $\emptyset \subset \cdots \subseteq [\varphi^{i+1}] \subseteq [\varphi^i] \subseteq \cdots \subseteq [\varphi^0]$. Since a formula has a finite number of models, it falls out immediately from this that there must be a point at which further revision by φ does not change anything.

Corollary 4.1

If a revision operator $\circ$ satisfies postulates R_1 and R_8, then $\circ$ is stable.

Unfortunately, postulates R_{11-12} do not guarantee stability. Since these postulates require only that the agent places the models of $\overline{\varphi}$ as the least plausible interpretations, it becomes possible that an agent's plausibility ranking does not hold on to a core set of interpretations through successive revisions by φ.

Example 4.7: Stability

For the set of atoms $A = \{a, b\}$ consider an agent whose prior information is represented by the formula $\varphi = \neg b$, revises their beliefs with an operator $\circ$ that satisfies postulates R_{11-12}, and who ranks outcomes as shown in Figure 4.4. We have that

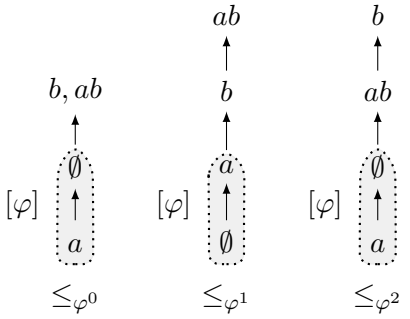


Figure 4.4: For an agent with prior information φ, $[\varphi] = \{\emptyset, a\}$, repeated revision by φ cycles between $\{a\}$ and $\{\emptyset\}$.

$[\varphi^0] = [\varphi] = \{\emptyset, a\}$, $[\varphi^1] = [\varphi \circ \varphi] = \{a\}$, and $[\varphi^2] = [\varphi^1 \circ \varphi] = \{\emptyset\}$. By postulate R_4, we infer that subsequent revisions by φ cycle between $\{a\}$ and $\{\emptyset\}$, i.e., $[\varphi^3] = \{a\}$, $[\varphi^4] = \{\emptyset\}$, and so on, therefore thus never settling on a stable answer.

The issue of stability suggests another dimension along which revision operators can be analyzed, with Corollary 4.1 and Example 4.7 showing that a revision operator does not satisfy it trivially. Example 4.7, in particular, shows that there is interplay between $\leq_\varphi$ and $\leq_{\varphi'}$, if $\varphi' \models \varphi$, which is relevant to the question of whether an operator is stable. This interplay is reminiscent of topics like iterated revision and kinetic consistency [Darwiche and Pearl, 1997, Peppas and Williams, 2016], but pursuing it further would take us too far afield of the aims of the current work.

4.4 Distance-based biased revision operators

Having characterized revision operators in terms of assignments on interpretations, we now ask: what is a natural way to construct operators with such biases? We will use the insight afforded by Theorem 4.1 to generate rankings on outcomes that reflect the design principles outlined by postulates R_{9-13}. In the process, we employ the two familiar ingredients from Section 2.3. The first is a quasi-distance d between interpretations, interpreted as a measure of plausibility of one interpretation relative to the other. The second ingredient is an aggregation function $\oplus$, used to compare interpretations given the distances generated by d. Putting these two ingredients together, we have a total $(d, \oplus)$-induced $\mathcal{L}$-assignment $\preccurlyeq_\varphi^{d,\oplus}$, which in turn induces the revision operator $\circ^{d,\oplus}$.

For quasi-distances, we will use drastic distance d_D and Hamming distance d_H. For aggregation functions, we use the ones introduced in Section 2.3, plus two new ones that we introduce in the following. The *d-centrality* $d^{\text{cen}}(\varphi, w)$ *of* w *with respect to* φ is defined as $d^{\text{cen}}(\varphi, w) = d^{\max}(\varphi, w) - d^{\min}(\varphi, w)$. The *d-displacement* $d^{\text{dis}}(\varphi, w)$ *of* w *with respect to* φ is $d^{\text{dis}}(\varphi, w) = d^{\min}(\varphi, w) - d^{\min}(w^*, \varphi)$, where w^* is an interpretation such that

$d^{\min}(w^*, \varphi)$ is minimal among all the interpretations w' for which $d^{\text{cen}}(w', \varphi) = d^{\text{cen}}(\varphi, w)$. Finally, the *d-agreeability* $d^{\text{agr}}(\varphi, w)$ *of* w *with respect to* φ is defined as $d^{\text{agr}}(\varphi, w) = \min\{d^{\min}(\varphi, w),\ d^{\text{cen}}(\varphi, w) + d^{\text{dis}}(\varphi, w)\}$, while the *d-disagreeability* $d^{\text{dagr}}(\varphi, w)$ *of* w *with respect to* φ is defined as $d^{\text{dagr}}(\varphi, w) = m - d^{\text{agr}}(\overline{\varphi}, w)$, where $m = |A|$.

Example 4.8: Aggregation functions

Take the formula $\varphi = (b \to a)$, with $[\varphi] = \{\emptyset, a, ab\}$, and the interpretation $w = \emptyset$. The vector of Hamming distances from every model of φ to w is:

$$(d_{\text{H}}(\emptyset, \emptyset), d_{\text{H}}(a, \emptyset), d_{\text{H}}(ab, \emptyset)) = (0, 1, 2).$$

We obtain that $d_{\text{H}}^{\text{leximax}}(\varphi, w) = (2, 1, 0)$ and $d_{\text{H}}^{\text{leximin}}(\varphi, w) = (0, 1, 2)$. Additionally, we have that:

$$\begin{aligned} d_{\text{H}}^{\min}(\varphi, w) &= 0, \\ d_{\text{H}}^{\max}(\varphi, w) &= 2, \\ d_{\text{H}}^{\text{sum}}(\varphi, w) &= 0 + 1 + 2 = 3, \\ d_{\text{H}}^{\text{cen}}(\varphi, w) &= 2 - 0 = 2, \\ d_{\text{H}}^{\text{agr}}(\varphi, w) &= (2 - 0) \cdot 0 = 0. \end{aligned}$$

Notice that the centrality of w with respect to φ is 0 just in case $d^{\min}(\varphi, w) = d^{\max}(\varphi, w)$, i.e., just in case w is at equal distance to every model of φ. Thus, the agreeability index of w with respect to φ is 0 just in case w is either a model of φ, or equally distanced to every model of φ.

Putting these ingredients together gives us the revision operators $\circ^{d,\,\min}$, $\circ^{d,\,\text{leximin}}$, $\circ^{d,\,\max}$, $\circ^{d,\,\text{leximax}}$, $\circ^{d,\,\text{sum}}$, $\circ^{d,\,\text{agr}}$, $\circ^{d,\,\text{dagr}}$, for $d \in \{d_{\text{H}}, d_{\text{D}}\}$. Out of these, $\circ^{\text{H},\,\min}$ is the Dalal operator and $\circ^{\text{D},\,\min}$ is the drastic operator, presented in Section 3.1. Thus, this perspective manages to capture known operators, while paving the way for some new ones.

The best way to understand these operators is to see how they rank interpretations in the universe.

Example 4.9: Distance-based biased preorders

For the set of atoms $A = \{a, b, c\}$, take $\varphi = (\neg(a \wedge b) \wedge \neg c) \vee (a \wedge b \wedge c)$, for which it holds that $[\varphi] = \{\emptyset, a, b, abc\}$. For the interpretation $w = \emptyset$, we have that $d_{\text{H}}^{\text{leximin}}(\varphi, w) = (0, 1, 1, 3)$, $d_{\text{H}}^{\text{leximax}}(\varphi, w) = (3, 1, 1, 0)$, $d_{\text{H}}^{\min}(\varphi, w) = 0$, $d_{\text{H}}^{\max}(\varphi, w) = 3$ and $d_{\text{H}}^{\text{sum}}(\varphi, w) = 5$. The distances and aggregated distances for each interpretation are depicted in Table 4.1. Notice how the models of φ are distributed when the interpretations are ranked according to the different aggregation functions used: we have $\emptyset \approx_{\varphi}^{\text{H},\,\min} a$, since $d_{\text{H}}^{\min}(\varphi, \emptyset) = d_{\text{H}}^{\min}(\varphi, a) = 0$, but $\emptyset <_{\varphi}^{\text{H},\,\text{leximin}} a$, since

	$\emptyset$	a	b	abc	leximin	leximax	min	max	sum
$\emptyset$	0	1	1	3	$(0,1,1,3)$	$(3,1,1,0)$	0	3	5
a	1	0	2	2	$(0,1,2,2)$	$(2,2,1,0)$	0	2	5
b	1	2	0	2	$(0,1,2,2)$	$(2,2,1,0)$	0	2	5
c	1	2	2	2	$(1,2,2,2)$	$(2,2,2,1)$	1	2	7
ab	2	1	1	1	$(1,1,1,2)$	$(2,1,1,1)$	1	2	5
ac	2	1	3	1	$(1,1,2,3)$	$(3,2,1,1)$	1	3	7
bc	2	3	1	1	$(1,1,2,3)$	$(3,2,1,1)$	1	3	7
abc	3	2	2	0	$(0,2,2,3)$	$(3,2,2,0)$	0	3	7

Table 4.1: The table of Hamming distances from the models of φ, with $[\varphi] = \{\emptyset, a, b, abc\}$, to every interpretation in a universe generated from three atoms. The aggregated values according to the aggregation functions presented in this chapter are also displayed.

$(0,1,1,3) <_{\text{lex}} (0,1,2,2)$. Also, we have that $c <_{\varphi}^{\text{H, max}} abc$, $c <_{\varphi}^{\text{H, leximax}} abc$ and $ab <_{\varphi}^{\text{H, sum}} abc$, i.e., models of φ are not minimal in $\leq_{\varphi}^{\text{H, max}}$, $\leq_{\varphi}^{\text{H, leximax}}$ and $\leq_{\varphi}^{\text{H, sum}}$. In particular, $\leq_{\varphi}^{\text{H, max}}$ makes the models of $\overline{\varphi}$ (i.e., abc, bc, ac and $\emptyset$) the least plausible interpretations.

The agreement and disagreement operators ($\circ^{d,\text{ agr}}$ and $\circ^{d,\text{ dagr}}$) are simpler than they appear: the idea behind $\circ^{d,\text{ agr}}$ is to allow interpretations other than the models of φ as the minimal elements of the preorder $\leq_{\varphi}$. Notice that the score of an interpretation in $\leq_{\varphi}^{d,\text{ agr}}$ is 0 if it is either a model of φ, or it is equidistant from every model of φ (i.e., its centrality is 0) and it is the 'closest' interpretation to φ with this property. The disagreement operator $\circ^{d,\text{ dagr}}$ works in similar fashion, by making models of $\overline{\varphi}$ and interpretations minimally equidistant to them the least plausible interpretations in $\leq_{\varphi}^{d,\text{ dagr}}$.

Example 4.10: Agreement and disagreement operators

If $A = \{a, b, c\}$, take φ such that $[\varphi] = \{a, b, c\}$, and notice that $\emptyset$ and abc are both equidistant to φ, hence their centrality is 0. However, $\emptyset$ is closer to φ than abc (its displacement is 0, compared to abc's displacement of 1), and $d_{\text{H}}^{\text{agr}}(\varphi, \emptyset) = 0$. Thus, what $d_{\text{H}}^{\text{agr}}$ does is to give a minimal score to models of φ and to the minimally equidistant interpretation $\emptyset$. By contrast, $d_{\text{H}}^{\text{dagr}}$ gives a maximal score to the models of $\overline{\varphi}$ and to the maximally equidistant interpretation abc.

All operators proposed generate a total preorder $\leq_{\varphi}^{d,\oplus}$ over interpretations, but differ in how they arrange models of φ: this corresponds to the different attitudes an agent can have towards φ prior to any revision. The operator $\circ^{\text{H, min}}$, known as Dalal's operator [Dalal, 1988], considers all models of φ as the most plausible elements in $\leq_{\varphi}^{\text{H, min}}$

and is the only operator of the ones considered here for which $\varphi \circ \mu$ is equivalent to $\varphi \wedge \mu$ when $\varphi \wedge \mu$ is consistent. Similarly, $\circ^{\mathrm{H},\,\mathrm{leximin}}$ also ranks models of φ as more plausible than any other interpretation, but discriminates among models of φ according to how typical, or representative they are of the general point of view expressed by φ. The operators $\leq_\varphi^{\mathrm{H},\,\max}$ and $\leq_\varphi^{\mathrm{H},\,\mathrm{leximax}}$ push away models of $\overline{\varphi}$, under the assumption that they are the most implausible possible worlds. They difference between them is that $\leq_\varphi^{\mathrm{H},\,\max}$ considers models of $\overline{\varphi}$ equally implausible, whereas $\leq_\varphi^{\mathrm{H},\,\mathrm{leximax}}$ uses the more fine-grained lexicographic approach. The operator $\circ^{\mathrm{H},\,\mathrm{agr}}$ makes models of φ the most plausible elements in $\leq_\varphi$ but does not stop here and also allows other interpretations on that position, in particular certain interpretations that are equidistant to φ as per Example 4.10. Specifically, an interpretation can be on the lowest level of $\leq_\varphi$ if it either is a model of φ, or is at equal distance to every model of φ. The intuition is that an interpretation equally distanced from models of φ is like a compromise point of view, with good chances of being correct if it is close to φ. The operator $\circ^{\mathrm{H},\,\mathrm{dagr}}$ is the dual of $\circ^{\mathrm{H},\,\mathrm{agr}}$ and, finally, operator $\circ^{\mathrm{H},\,\mathrm{sum}}$ evokes utilitarian approaches by choosing interpretations that minimize the sum of the distances to each model of φ, i.e., are close to φ on an aggregate level.

Plugging in the drastic and Hamming distances would seem to give us a considerable number of operators, but quick reflection shows that operators obtained with drastic distance d_{D} collapse into two main categories. To get a grasp on this fact, consider first the *drastic revision operator* $\circ^{\mathrm{D}}$ defined, for any propositional formulas φ and μ, as $\varphi \circ^{\mathrm{D}} \mu = \varphi \wedge \mu$, if $\varphi \wedge \mu$ is consistent, and μ otherwise, and the *forgetful revision operator* $\circ^{\mathrm{F}}$ defined as $\varphi \circ^{\mathrm{F}} \mu = \mu$. It is forgetful because it disregards initially held beliefs completely, always adopting the new information μ.

Proposition 4.2

For any propositional formulas φ and μ, it holds that $\varphi \circ^{\mathrm{D},\,\min} \mu \equiv \varphi \circ^{\mathrm{D},\,\mathrm{leximin}} \mu \equiv \varphi \circ^{\mathrm{D},\,\mathrm{leximax}} \mu \equiv \varphi \circ^{\mathrm{D},\,\mathrm{sum}} \mu \equiv \varphi \circ^{\mathrm{D}} \mu$. Moreover, $\varphi \circ^{\mathrm{D},\,\mathrm{agr}} \mu \equiv \varphi \circ^{\mathrm{F}} \mu$ and $\varphi \circ^{\mathrm{D},\,\max} \mu \equiv \varphi \circ^{\mathrm{D},\,\mathrm{dagr}} \mu \equiv \begin{cases} \varphi \circ^{\mathrm{D}} \mu, & \text{if } \varphi \text{ is complete,} \\ \varphi \circ^{\mathrm{F}} \mu, & \text{otherwise.} \end{cases}$

Proof

If φ is a complete formula such that $[\varphi] = \{v\}$, then $d_{\mathrm{D}}^{\oplus}(\varphi, w) = d_{\mathrm{D}}(v, w)$, for all aggregation functions $\oplus \in \{\min, \max, \mathrm{leximin}, \mathrm{leximax}, \mathrm{sum}, \mathrm{agr}, \mathrm{dagr}\}$, and any interpretation w. In other words, it holds that:

$$d_{\mathrm{D}}^{\oplus}(\varphi, w) = \begin{cases} 0, & \text{if } w = v, \\ 1, & \text{otherwise.} \end{cases}$$

It is thus straightforward to conclude that if $v \in [\mu]$, then $[\varphi \circ^{\mathrm{D},\,\max} \mu] = \{v\} = [\varphi \wedge \mu]$,

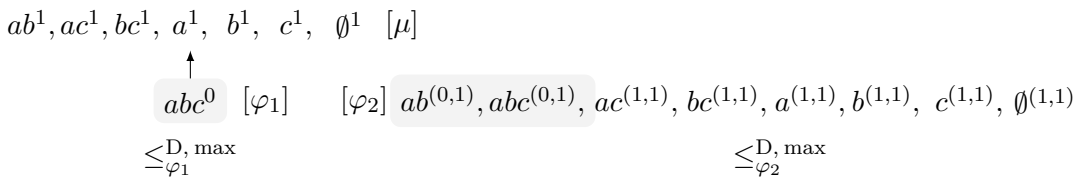


Figure 4.5: Preorders $\leq_{\varphi_1}^{\mathrm{D,\,max}}$ and $\leq_{\varphi_2}^{\mathrm{D,\,max}}$, for $[\varphi_1] = \{abc\}$ and $[\varphi_2] = \{ab, ac, abc\}$. Depicted as superscript to an interpretation w is the vector of drastic distances d_{D} from the models of φ_1 and φ_2, respectively, to w. Since φ_1 is complete (i.e., has only one model), this vector consists of only one number. The max aggregation function uses the maximum value in the vector of distances to compare interpretations.

and if $w_0 \notin [\mu]$, then $[\varphi \circ^{\mathrm{D,\,max}} \mu] = [\mu]$. If φ is not complete, then we have that:

$$d_{\mathrm{D}}^{\mathrm{leximin}}(\varphi, w) = \begin{cases} (0, 1, \ldots, 1), & \text{if } w \in [\varphi], \\ (1, 1, \ldots, 1), & \text{otherwise.} \end{cases}$$

It is now straightforward to see that the remaining statements of Proposition 4.2 hold.

We can illustrate the differential treatment of operator $\circ^{\mathrm{D,\,max}}$ through an example.

Example 4.11: Drastic operators

For the set of atoms $A = \{a, b, c\}$, consider the complete formula $\varphi_1 = \{a \wedge b \wedge c\}$. We have that:

$$\begin{aligned} d_{\mathrm{D}}^{\max}(\varphi, w) &= d_{\mathrm{D}}(abc, w) \\ &= \begin{cases} 0, & \text{if } w = abc, \\ 1, & \text{otherwise.} \end{cases} \end{aligned}$$

The preorder $\leq_{\varphi_1}^{\mathrm{D,\,max}}$ is depicted on the left in Figure 4.5. Consider, then, a formula $\varphi_2 = a \wedge b$, with $[\varphi_2] = \{ab, abc\}$. In this case we obtain, for instance, that $d_{\mathrm{D}}^{\max}(\varphi, w)$, for every interpretation w in the universe. The preorder $\leq_{\varphi_2}^{\mathrm{D,\,max}}$ is depicted on the right in Figure 4.5, which illustrates the fact that in this case $\circ^{\mathrm{D,\,max}}$ is equivalent to the forgetful operator $\circ^{\mathrm{F}}$.

With Hamming distance the landscape is more diverse, as the different attitudes the operators assume towards models of φ lead to genuinely different revision strategies. Nonetheless, certain relationships between the operators still hold, with lexicographic operators being the most discriminating, in the sense that they pick formulas with fewer models, i.e., more specific formulas.

	R_9	R_{10}	R_{11}	R_{12}	R_{13}	R_{NEUT}	R_{IAHB}	R_{STAB}
$\circ^{\mathrm{H,\,min}}$	✓	✓	×	×	✓	✓	✓	✓
$\circ^{\mathrm{H,\,leximin}}$	✓	×	×	×	✓	✓	×	✓
$\circ^{\mathrm{H,\,agr}}$	×	✓	×	×	✓	✓	✓	✓
$\circ^{\mathrm{H,\,max}}$	×	×	✓	✓	×	✓	×	✓
$\circ^{\mathrm{H,\,leximax}}$	×	×	×	✓	×	✓	×	✓
$\circ^{\mathrm{H,\,dagr}}$	×	×	✓	×	×	✓	✓	✓
$\circ^{\mathrm{H,\,sum}}$	×	×	×	×	✓	✓	×	✓
$\circ^{\mathrm{D}}$	✓	✓	×	×	✓	✓	✓	✓
$\circ^{\mathrm{F}}$	×	✓	✓	×	✓	✓	✓	✓

Table 4.2: Satisfaction of postulates for the biased operators $\circ^{d,\,\oplus}$ described in this chapter. We include rows for operators $\circ^{\mathrm{D}}$ and $\circ^{\mathrm{F}}$ rather than have separate rows for the operators generated using drastic distance d_{D}, with the understanding given by Proposition 4.2 that they collapse, in one way or another, into one of these two.

Proposition 4.3

If φ and μ are propositional formulas and d is a quasi-distance, it holds that:

(*a*) $\varphi \circ^{d,\,\mathrm{leximin}} \mu \models \varphi \circ^{d,\,\mathrm{min}} \mu \models \varphi \circ^{d,\,\mathrm{agr}} \mu$,

(*b*) $\varphi \circ^{d,\,\mathrm{leximax}} \mu \models \varphi \circ^{d,\,\mathrm{max}} \mu \models \varphi \circ^{d,\,\mathrm{dagr}} \mu$.

All operators generate total preorders over interpretations, so by Theorem 3.5 they all satisfy postulates R_1 and R_{3-6}. Satisfaction with respect to the newly introduced postulates is clarified below, but before we state the result we introduce a property of distances that will be useful in settling the matter with respect to neutrality. This property is expected to hold for any interpretations w_1 and w_2 and renaming ρ of the atoms in A:

($\mathrm{D}_{\mathsf{NEUT}}$) $d(w_1, w_2) = d(\rho(w_1), \rho(w_2))$.

A quasi-distance d is *neutral* if it satisfies property $\mathrm{D}_{\mathsf{NEUT}}$. With this in hand, we can introduce the result.

Proposition 4.4

For $d \in \{d_{\mathrm{D}}, d_{\mathrm{H}}\}$ and $\oplus \in \{\mathrm{min}, \mathrm{leximin}, \mathrm{max}, \mathrm{leximax}, \mathrm{agr}, \mathrm{dagr}, \mathrm{sum}\}$, the operators $\circ^{d,\,\oplus}$ satisfy postulates R_{9-13}, R_{NEUT}, R_{IAHB} and R_{STAB} as shown in Table 4.2.

Proof

We will use Theorem 4.1 to show that the operators arrange interpretations in the patters described by the properties r_{7-11}.

It is already known that $\leq_\varphi^{H,\,min}$, known as Dalal's operator [Dalal, 1988], satisfies postulate R_6, which implies that it satisfies postulates R_9 and R_{10}. To see why the operator $\circ^{H,\,leximin}$ satisfies R_9, notice that if $w_1 \in [\varphi]$ and $w_2 \notin [\varphi]$, then the first element in $d_H^{leximin}(\varphi, w_1)$ is 0, while the first element in $d_H^{leximin}(\varphi, w_2)$ is strictly greater than 0. This implies that $d_H^{leximin}(\varphi, w_1) <_{lex} d_H^{leximin}(\varphi, w_2)$, which in turn implies that $w_1 <_\varphi^{H,\,leximin} w_2$. Hence property r_7 is satisfied, which implies that postulate R_9 is satisfied. For $\circ^{H,\,leximin}$ and postulate R_{10}, take $[\varphi] = \{a, b, ab\}$ and $[\mu] = \{a, b, ab\}$. We get that $[\varphi \circ^{H,\,leximin} \mu] = \{a, b\}$ and thus $\varphi \wedge \mu \not\models \varphi \circ^{H,\,leximin} \mu$. The operator $\circ^{H,\,agr}$ satisfies postulate R_{10} because it makes all models of φ, and potentially other interpretations as well (which is the reason why it does not satisfy postulate R_9), as the equally most plausible interpretations in $\leq_\varphi^{H,\,agr}$. Since all these operators place the models of φ on the lowest levels of $\leq_\varphi$, they all satisfy postulate R_{13}.

To see why postulates R_{11} and R_{12} are not satisfied by $\circ^{H,\,\oplus}$, for $\oplus \in \{min, leximin, agr\}$, it is sufficient to notice that these operators do not make models of $\overline{\varphi}$ as the least plausible interpretations in $\leq_\varphi^{H,\,\oplus}$. Thus, if $\varphi = a \vee b$, then $\overline{\varphi}$ shares some models with φ, yet these models (along with all other models of $a \vee b$) will be among the most plausible interpretations in $\leq_\varphi^{H,\,min}$, $\leq_\varphi^{H,\,leximin}$ and $\leq_\varphi^{H,\,agr}$, due to how these preorders are defined. The one exception is the forgetful operator $\circ^F$, which satisfies R_{11} trivially.

The case for $\circ^{H,\,max}$, $\circ^{H,\,leximax}$ and $\circ^{H,\,dagr}$ is analogous to the one for $\circ^{H,\,min}$, $\circ^{H,\,leximin}$ and $\circ^{H,\,agr}$, as they can be seen as duals of each other. For the operator $\circ^{H,\,sum}$ and postulates R_{9-10}, take $[\varphi] = \{a, b, c\}$ and $[\mu] = \{\emptyset, a, b, c\}$. We get that $[\varphi \circ^{H,\,sum} \mu] = \{\emptyset\}$, as $\emptyset$ minimizes the sum of the Hamming distances to the models of φ. For postulates R_{11-12}, take $[\mu'] = \{\emptyset, ab, ac, bc\}$. For $\circ^{H,\,sum}$ and R_{13}, notice that adding v to $[\varphi]$ creates a new column for v in the table of distances, in which the distance corresponding to v is 0, i.e., the score assigned to w' in $\leq_{\varphi\vee\varphi'}^{H,\,sum}$ does not increase with respect to $\leq_\varphi^H$ sum. Satisfaction of R_{IAHB} and R_{STAB} is straightforward, keeping in mind how the various operators arrange the models of φ in the generated preorders.

For the neutrality postulate R_{NEUT}, it is straightforward to see that the drastic and Hamming distances are neutral. Furthermore, if d is neutral, then it follows straightforwardly that $d^\oplus(\varphi, w) = d^\rho(\rho(\varphi), \rho(w))$, for any $w \in \mathcal{U}$, and $w_1 \leq_\varphi^{d,\,\oplus} w_2$ iff $\rho(w_1) \leq_{\rho(\varphi)}^{d,\,\oplus} \rho(w_2)$, for all aggregation functions introduced so far. Thus, the preorders $\leq_\varphi^{d,\,\oplus}$ satisfy property r_{NEUT} and, by Theorem 4.1, the operators represented by them satisfy postulate R_{NEUT}.

It should be kept in mind that neutrality is not guaranteed by the standard postulates R_{1-6}, or by any of the other postulates introduced so far, but the way in which concrete operators are usually defined (i.e., by appeal to neutral distances) indicates that neutrality is part of our basic understanding of how a revision operator should behave. And, in general, there seems to be no a priori reason for looking at non-neutral operators. However, we will see in Chapter 6 that such operators cannot be avoided when we move to a fragment of propositional logic.

4.5 Related work

The idea that revision postulate R_2 sometimes leads to counterintuitive results has been remarked upon before.

Example 4.12: John

Consider the example of John, proposed by Fermé and Hansson:

> John is a neighbour about whom I initially know next to nothing.
>
> *Case 1*: I am told that he goes home from work by taxi every day (t). This makes me believe that he is a rich man (r).
>
> *Case 2*: When told t, I am also told that John is a driver by profession (d). In this case I am not made to believe that he is a rich man (r). [Fermé and Hansson, 2018, p. 45]

In our terminology, the set of atoms is $A = \{t, r, d\}$. The agent's initial beliefs are φ: we are not told what φ is, but we are led to believe that it is consistent with t, r and d. Let us assume that $\varphi = \top$. There are two instances of revision, once by $\mu_1 = t$, and then by $\mu_2 = t \wedge d$. We are told that $\varphi \circ \mu_1 \models r$, but $\varphi \circ \mu_2 \not\models r$. Assuming the agent is revising according to a total preorder $\leq_\varphi$ on interpretations, the claims above translate to $\min_{\leq_\varphi}\{t, tr, td, trd\} \subseteq \{r, tr, rd, trd\}$, but $\min_{\leq_\varphi}\{td, trd\} \not\subseteq \{r, tr, rd, trd\}$. This is not possible if the agent considers all interpretations as equally plausible: the agent must consider, for instance, that $tr <_\varphi td$ and $tr <_\varphi tdr$. This is clearly in conflict with postulate R_2.

Fermé and Hansson explain Example 4.12 by appeal to a type of cognitive attitude that an agent might plausibly adopt when revising:

> When we acquire a new belief that does not contradict our previous beliefs (such as t in the example), we often include in the outcome some additional belief (such as r in the example) that does not follow deductively but nevertheless serves to make the belief set more complete and/or more coherent. [Fermé and Hansson, 2018, p. 45]

Presumably, John jumps to conclusions according to some stereotypical images about rich people and cars. This is exactly the type of attitude exhibited in Examples 4.1 and 4.2. Later in the same section, Fermé and Hansson also present another example.

Example 4.13: Valentina

Consider the example of Valentina, proposed by Fermé and Hansson:

> Valentina was uncertain whether or not her husband is unfaithful to her (u), but she still believed that her husband loves her (l). However, when she learnt that he is unfaithful to her, she lost her belief that he loves her. [Fermé and Hansson, 2018, p. 45]

In our terminology, the set of atoms is $A = \{u, l\}$, Valentina's prior belief is $\varphi = l$, the new information is $\mu = u$, and Valentina's posterior information after revision is $\varphi \circ \mu \equiv u \wedge \neg l$, whereas postulate R_2 requires that $\varphi \circ \mu \equiv u \wedge l$.

Right after presenting this example Fermé and Hansson mention that postulate R_2, or, as they call it, the *expansion property of revision*, "has been much less discussed than the recovery property of contraction, but it is no less problematic and no less difficult to remove from the AGM framework" [Fermé and Hansson, 2018, p. 45].

A response to these kinds of examples has been the framework of *abductive expansion* [Pagnucco et al., 1994], in which addition of a new item of knowledge must come with a justification, or explanation for the new belief. A semantic model for this operation is sketched in terms of Groves' system of spheres [Grove, 1988], but not much more detail is added. Other than this, there are few other works considering revision operators that do not satisfy the classical postulate R_2 [Ryan, 1996, Benferhat et al., 2005]. As shown above, the acknowledgement that deviations from R_2 make sense is not entirely foreign, but the idea such deviations correspond to possible epistemic attitudes and can be realized through distance-based approaches has, to the best of our knowledge, not been considered in any systematic detail before.

4.6 Conclusion

In this chapter we have looked at the classical revision postulates from the point of view of what they assume about an agent's attitude towards its initial beliefs, and argued that this attitude is embedded in a specific postulate, i.e., the standard postulate R_2. By varying this postulate and calling attention to a commonly overlooked neutrality property, we were able to put forward and characterize a wide range of revision operators, and refine previously entangled intuitions in the process. We also showed that this level of analysis is needed when working in restricted fragments of propositional logic, where postulate R_2 cannot be satisfied and must therefore be broken down into two separate components (postulates R_{7-8} in the current work). The aggregation functions used to

construct revision operators recall methods to rank outcomes in decision theory. Analysis of the new operators also uncovered the principles of indifference to already held beliefs (R_{IAHB}) and stability (R_{STAB}). Further work is needed to link these notions to the other postulates, to map out their interplay and to provide them with semantic characterizations. Following the line of reasoning initiated in the previous section, a natural follow-up would be to consider the proposed postulates in fragments of propositional logic and to look for characterizations in terms of preorders on outcomes.

Discussions of stability and biases notwithstanding, one might still question the rationale behind doubting postulate R_2: indeed, why fix something if it is not broken? In response, we will see in Chapter 6 that there situations where revision is warranted, but in which postulates R_{9-10} *cannot* occur together. In Chapter 6 we will look at revision of Horn formulas: as mentioned in the introduction, there is good reason to want to do revision on such specialized formulas, and we will see that postulate R_2 is at odds with the expressibility requirements of such a revision operator.

CHAPTER 5

Merging as Fair Collective Choice

In this chapter we look at merging as a collective decision mechanism akin to an election, whose goal is to aggregate information originating with different agents. As mentioned in Section 3.4, our approach in this work sees merging as a task whose role is not so much to find the true answer, but rather to find a compromise between the different opinions of the participans. In this, our main purpose is to make sure that the aggregation process is *fair* towards the agents involved, in all the various ways that fairness can be conceived of: to this end, we look to the social choice literature, which contains an arsenal of properties that have been used to ensure fairness of voting rules [Zwicker, 2016, Baumeister and Rothe, 2016], and seek to apply these properties to the context of merging. This involves, first of all, refitting the main intuitions to the context of merging, which is not always straightforward, and seeing to what extent existing merging operators satisfy the newly minted properties. In some cases, we take cues from the social choice literature even to design new merging operators, tailored specifically to these properties.

What makes the appropriation of classical social choice properties challenging, in certain cases, are the differences between merging operators and classical voting rules. Though merging operators can be seen as social choice functions, as mentioned in Section 3.4, they differ from standard voting rules as analyzed in social choice theory, in that they do not require agents to rank all possible alternatives. What agents provide to a merging operator are formulas: if we identify formulas with their models, and think of the models as candidates up for election, then, under postulate M_2, merging operators can be seen as social choice functions that require agents to submit only their top preferences. Nonetheless, the representation result in Section 3.4 shows how, under certain assumptions, preferences creep in even when not explicitly provided.

Example 5.1: #OscarsSoFossilized, again

We return to the four Academy members from Example 1.5 trying to decide the nominees for the category of Best Director. The three options (Alma Har'el, Bong Joon Ho and Céline Sciamma) are represented by propositional atoms a, b and c. The opinions of the four Academy members are represented by the formulas $\varphi_1 = a \wedge b$, $\varphi_2 = a \wedge (b \vee c)$, $\varphi_3 = \neg a \wedge b \wedge \neg c$. and $\varphi_4 = \neg a \wedge \neg b \wedge c$, gathered in the profile $\vec{\varphi} = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$. The list has to be whittled down to two, i.e., there is a constraint $\mu = (a \wedge b \wedge \neg c) \vee (a \wedge \neg b \wedge c) \vee (\neg a \wedge b \wedge c)$, with $[\mu] = \{ab, bc, ac\}$, that needs to be satisfied. A merging operator Δ satisfying postulates M_{0-8} delivers a propositional formula $\Delta_\mu(\vec{\varphi})$ that, among other things, satisfies μ. What is more, according to Theorem 3.10 we know that every formula $\varphi_i \in \vec{\varphi}$ induces a total preorder $\leq_{\varphi_i}$ on outcomes.

Thus, merging the formulas in $\vec{\varphi}$ can be seen as an election where the voters are the Academy members (i.e., the agents supplying the formulas), the candidates are the viable nominee lineups (i.e., the models of μ) and the voting rule is the merging operator Δ. In this context, the agents' beliefs can be seen as encoding their top options: thus, Academy member 1's opinion φ_1 has as models the interpretations ab and abc, which, according to postulate M_2, are the Academy member's most preferred outcomes in their corresponding preference order $\leq_{\varphi_1}$. In Section 3.4 we have also seen that distances between interpretations and aggregation functions can be used to generate a total preorder based on the opinions provided by the agents, i.e., it need not be assumed that agents hold the preference order is their 'heads', or have to explicitly provide them.

Thus, though belief merging operators and voting rules share a common goal and methodology, the parameters of a belief merging operator are subtly different from those of a classical voting rule, with the closest match in social choice being models of combinatorial voting based on completion principles [Lang and Xia, 2016]. Nonetheless, we believe that the wealth of insights accumulated by social choice theory on voting methods can be brought to bear on merging operators.

The main thrust of this chapter lies in a series of properties meant to capture various aspects of fairness in the merging process. We present these properties as postulates that a merging operator Δ is expected to satisfy, and intend them as additions to the standard merging postulates: our purpose, to be clear, is not to suggest that postulates M_{0-8} are wrong, or that they have to be replaced; the postulates we look at are meant to stand alongside the existing postulates and complement them. As such, our contribution aims to extend the standard characterization of merging operators by offering more fine-grained criteria for their evaluation. We group the properties according to their character, and offer discussions on the behavior they are intended to model. In the case of each new property, we study its relationship with the core postulates M_{0-8}. When a property is not guaranteed by these postulates, we investigate which of the standard

operators satisfy the property, give relevant counter-examples, and provide model-based representation results for the most prominent of these properties.

5.1 Insensitivity to syntax

The properties in this section are grouped around the idea that the outcome of a merging task should depend only on the semantic content of the information provided and not on details about how the information is written down, perceived here as extraneous. More concretely, the idea is that aspects of the syntax of the elements of merging should not affect the result of the merging process. This is an intuition that is already familiar to us, since the standard postulate M_3 already expresses a form of insensitivity to syntax. However, there are more nuances to this principle than M_3 manages to capture.

Before presenting the actual properties we have in mind, we must become acquainted with some notions, some of them new and some of them old. The first notions describe ways of swapping things around in a profile. Recall that a permutation σ of the set of agents $N = \{1, \dots, n\}$ is a bijection $\sigma\colon N \to N$. If $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ is a propositional profile and σ is a permutation of N, the *permutation σ of $\vec{\varphi}$* is defined as the profile $\vec{\varphi} = (\varphi_{\sigma(i)})_{1 \leq i \leq n}$, i.e., the profile obtained by changing the order of the formulas in $\vec{\varphi}$ in accordance with σ. A renaming ρ of the set A of atoms is a permutation of the atoms in A, and is familiar from Section 4.1. We extend it now to profiles, and say that if ρ is a renaming of A and $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ is a propositional profile, *the renaming $\rho(\vec{\varphi})$ of $\vec{\varphi}$* is the profile $\vec{\varphi} = (\rho(\varphi_i))_{1 \leq i \leq n}$.

The next notion describe ways of getting rid of certain types of information, which, for some reason, may become redundant. If p and q are atoms in A and φ is a propositional formula, the *bundling $\varphi^{p \rightsquigarrow q}$ of p into q in φ* is the formula obtained by replacing every occurrence of p in φ_i with q. If $\vec{\varphi} = (\varphi_1, \dots, \varphi_n)$ is a propositional profile, the *bundling $\vec{\varphi}^{p \rightsquigarrow q}$ of p into q in $\vec{\varphi}$* is the profile $\vec{\varphi}^{p \rightsquigarrow q} = (\varphi_1^{p \rightsquigarrow q}, \dots, \varphi_n^{p \rightsquigarrow q})$, obtained by replacing every occurrence of p in φ_i with q, for every $\varphi \in \vec{\varphi}$.

We can now introduce the actual postulates. They are intended to apply for any profile $\vec{\varphi}$, propositional formula μ, permutation σ of $\mathbb{N}$, renaming ρ of the atoms in A and atoms p and q:

($\mathrm{M}_{\mathsf{ANON}}$) $\Delta_\mu(\vec{\varphi}) \equiv \Delta_\mu(\sigma(\vec{\varphi}))$.

($\mathrm{M}_{\mathsf{NEUT}}$) $\rho(\Delta_\mu(\vec{\varphi})) \equiv \Delta_{\rho(\mu)}(\rho(\vec{\varphi}))$.

($\mathrm{M}_{\mathsf{BNDL}}$) $\Delta_{\mu \wedge (p \leftrightarrow q)}(\vec{\varphi}) \models \Delta_{\mu^{p \rightsquigarrow q}}(\vec{\varphi}^{p \rightsquigarrow q})$.

Postulate $\mathrm{M}_{\mathsf{ANON}}$, where '$\mathsf{ANON}$' stands for *anonymity*, requires that the result of a merging task is invariant under permutations of the formulas in a profile, and is an analogue of the anonymity property often encountered in voting [Baumeister and Rothe, 2016, Zwicker, 2016]. It is a desirable property if, as is usually the case, it is felt that the

result should not depend on the order in which agents submit their opinions: it should not matter to the aggregation process what an agent's social security number is, or at what time in the day the agent submits its preferences. An $\mathcal{L}^n$-merging operator Δ is *anonymous* if it satisfies postulates $\mathrm{M}_{\mathsf{ANON}}$.

Postulate $\mathrm{M}_{\mathsf{NEUT}}$, where 'NEUT' stands for *neutrality*, says that the result of a merging task is invariant under permutations of the atoms. Postulate $\mathrm{M}_{\mathsf{NEUT}}$ is a close analogue of the revision postulate $\mathrm{R}_{\mathsf{NEUT}}$ from Section 4.1. In a voting scenario neutrality requires that if two candidates are swapped in all votes, then they are also swapped in the result. Its purpose is to ensure that all candidates are treated equally in the determination of the winners, i.e., their names do not matter. Since in the context of merging candidates are outcomes and outcomes are uniquely identified by the atoms that are true in them, we apply the neutrality property at the level of the atoms. An $\mathcal{L}^n$-merging operator Δ is *neutral* if it satisfies postulates $\mathrm{M}_{\mathsf{NEUT}}$.

The last postulate in this section is $\mathrm{M}_{\mathsf{BNDL}}$, where 'BNDL' stands for *bundling*. Though it has no direct equivalent in the voting literature, postulate $\mathrm{M}_{\mathsf{BNDL}}$ bears some resemblance to a property of voting rules called *Independence of clones* [Baumeister and Rothe, 2016], and is motivated by a similar intuition: alternatives that are in some sense redundant should not skew the vote in their favour. The intuition behind this property is best illustrated by an example.

Example 5.2: Bundling

In the scenario described in Example 5.1, with four Academy members trying to decide who will be the *Best Director* nominees, a very unlikely thing happens: Alma Har'el and Bong Joon Ho are discovered to be the same person. The show must go on, of course, but unfortunately this revelation comes after the Academy members have submitted their opinions, and it is too late to go back and have them redo their evaluation. What is known for sure, however, is that any distinction between Alma Har'el and Bong Joon Ho in the decision process has to be erased.

Example 5.2 provides a motivation for the bundling postulate $\mathrm{M}_{\mathsf{BNDL}}$: at some point in the modeling process, variables p and q, which hitherto had been thought to stand for different things, are discovered to encode the same concept. One way to incorporate this information in the merging process is by 'bundling' p into q in the formulas and in the constraint: as it were, cutting every occurrence of p and pasting q where p had been. This is quite a laborious and invasive operation on the formulas, which might be infeasible if access to the formulas is limited or if the formulas are provided by the agents just in time for the merging process. Another way is to add the equivalence $p \leftrightarrow q$ directly to the constraint and enforce that p and q are tied up together in perpetuity. Postulate $\mathrm{M}_{\mathsf{BNDL}}$ explores the relationship between these two operations and requires that all solutions of the latter operation are also solutions of the former.

Example 5.3: Anonymity, neutrality and bundling

For the profile $\vec{\varphi} = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$ with $\varphi_1 = a \wedge b$, $\varphi_2 = a \wedge (b \vee c)$, $\varphi_3 = \neg a \wedge b \wedge \neg c$ and $\varphi_4 = \neg a \wedge \neg b \wedge c$, the constraint $\mu = (a \wedge b \wedge \neg c) \vee (a \wedge \neg b \wedge c) \vee (\neg a \wedge b \wedge c)$, the merging operator $\Delta^{\mathrm{H},\,\min}$ and a permutation σ such that $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 4$ and $\sigma(4) = 1$, we obtain that $[\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi})] = \{ab, bc\}$ and:

$$\begin{aligned}[\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi})] &= [\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\varphi_1, \varphi_2, \varphi_3, \varphi_4)] \\ &= [\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\varphi_2, \varphi_3, \varphi_4, \varphi_1)] \\ &= [\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\varphi_{\sigma(1)}, \varphi_{\sigma(2)}, \varphi_{\sigma(3)}, \varphi_{\sigma(4)})] = [\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\sigma(\vec{\varphi}))].\end{aligned}$$

This is consistent with postulate $\mathrm{M}_{\mathsf{ANON}}$.

If ρ is a renaming of A such that $\rho(a) = b$, $\rho(b) = c$ and $\rho(c) = a$, we have that $\rho(\varphi_1) = b \wedge c$, $\rho(\varphi_2) = b \wedge (c \vee a)$, $\rho(\varphi_3) = \neg b \wedge c \wedge \neg a$, $\rho(\varphi_4) = \neg b \wedge \neg c \wedge a$, $\rho(\mu) \equiv \mu$, and $[\Delta_{\rho(\mu)}^{\mathrm{H},\,\mathrm{sum}}(\rho(\vec{\varphi}))] = \{bc, ca\} = [\rho(\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}))]$, which is consistent with postulate $\mathrm{M}_{\mathsf{NEUT}}$.

If we bundle b into a, we have that $\varphi_1^{b \rightsquigarrow a} = a \wedge a$, $\varphi_2^{b \rightsquigarrow b} = a \wedge (a \vee c)$, $\varphi_3^{b \rightsquigarrow a} = \neg a \wedge a \wedge \neg c$, $\varphi_4^{b \rightsquigarrow a} = \neg a \wedge \neg a \wedge c$ and $\mu^{b \rightsquigarrow a} = (a \wedge a \wedge \neg c) \vee (a \wedge \neg a \wedge c) \vee (\neg a \wedge a \wedge c)$. We obtain that $[\Delta_{\mu^{b \rightsquigarrow a}}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{b \rightsquigarrow a})] = \{ab\}$. On the other hand, note that $[\mu \wedge (a \leftrightarrow b)] = \{ab\}$ and thus $[\Delta_{\mu \wedge (a \leftrightarrow b)}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi})] = \{ab\}$. This result is consistent with postulate $\mathrm{M}_{\mathsf{BNDL}}$.

To understand the postulates just introduced it is, of course, useful to see how they work on the semantic level, i.e., to understand what kind of properties need to hold for the preorders in an $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations that represents the merging operator. Thus, if $\preccurlyeq$ is an $\mathcal{L}^n$-assignment on interpretations that satisfies properties m_{1-9} and ρ is a renaming on A, the following properties turn out to be relevant, when applying for any propositional profile $\vec{\varphi}$ and interpretations w_1 and w_2:

($\mathrm{m}_{\mathsf{ANON}}$) If $w_1 \leq_{\vec{\varphi}} w_2$, then $w_1 \leq_{\sigma(\vec{\varphi})} w_2$.

($\mathrm{m}_{\mathsf{NEUT}}$) If $w_1 \leq_{\vec{\varphi}} w_2$, then $\rho(w_1) \leq_{\rho(\vec{\varphi})} \rho(w_2)$.

Property $\mathrm{m}_{\mathsf{ANON}}$ says that if w_1 is considered at least as good as w_2 in the preorder corresponding to $\vec{\varphi}$, then this situation should be preserved in the preorder corresponding to $\sigma(\vec{\varphi})$. Since the inverse σ^{-1} of σ is also a permutation, property $\mathrm{m}_{\mathsf{ANON}}$ implies, of course, that $\leq_{\vec{\varphi}} = \leq_{\sigma(\vec{\varphi})}$, i.e., the preorder $\leq_{\vec{\varphi}}$ assigned to profile $\vec{\varphi}$ is identical to the preorder $\leq_{\sigma(\vec{\varphi})}$ assigned to the profile $\sigma(\vec{\varphi})$, obtained by permuting the formulas in $\vec{\varphi}$ according to σ. Property $\mathrm{m}_{\mathsf{NEUT}}$ expresses a similar property, but with respect to renamings, and likewise implies that $\leq_{\vec{\varphi}} = \leq_{\rho(\vec{\varphi})}$, i.e., the preorder $\leq_{\vec{\varphi}}$ associated to a profile $\vec{\varphi}$ is identical to the preorder $\leq_{\rho(\vec{\varphi})}$ assigned to the profile $\rho(\vec{\varphi})$, obtained by

renaming $\vec{\varphi}$. An $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations is *anonymous* if it satisfies property $\mathrm{m}_{\mathsf{ANON}}$ and *neutral* if it satisfies property $\mathrm{m}_{\mathsf{NEUT}}$.

Intuitively, we would expect that postulates $\mathrm{M}_{\mathsf{ANON}}$ and $\mathrm{M}_{\mathsf{NEUT}}$ map onto properties $\mathrm{m}_{\mathsf{ANON}}$ and $\mathrm{m}_{\mathsf{NEUT}}$, i.e., that anonymous and neutral merging operators are characterized by anonymous and neutral assignments, respectively. And indeed, this is what we find. For the next result, recall that an $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations represents an $\mathcal{L}^n$-merging operator Δ if, for any propositional profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$ and propositional formula μ, it holds that $[\Delta_\mu(\vec{\varphi})] = \min_{\leq_{\vec{\varphi}}}[\mu]$.

Theorem 5.1

If Δ is an $\mathcal{L}^n$-merging operator that satisfies postulates M_{0-1} and M_3, and $\preccurlyeq$ is a total $\mathcal{L}^n$-assignment on interpretations that represents it, then the following equivalences hold:

(1) Δ satisfies postulate $\mathrm{M}_{\mathsf{ANON}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{ANON}}$.

(2) Δ satisfies postulate $\mathrm{M}_{\mathsf{NEUT}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{NEUT}}$.

Proof

Since for this proof we will use $\mathcal{L}$-proxies of a pair of interpretations $\{w_1, w_2\}$, and these are not necessarily unique, postulate M_4 is used to ensure that the results of a merging operator is invariant under any choice of $\mathcal{L}$-proxy. Postulates M_{0-1} are needed to ensure the existence of a well defined assignment that can represent Δ.

For Equivalence (1), note that $\vec{\varphi} \equiv \sigma(\vec{\varphi})$, for any permutation σ. Thus, if Δ satisfies postulate $\mathrm{M}_{\mathsf{ANON}}$ and $\preccurlyeq$ is a total $\mathcal{L}^n$-assignment on interpretations that represents it, then, for any two interpretations w_1 and w_2, profile $\vec{\varphi}$ and permutation σ, it holds that:

$$\begin{aligned} w_1 \leq_{\vec{\varphi}} w_2 \text{ iff } & w_1 \in [\Delta_{\varepsilon_{1,2}}(\vec{\varphi})] && \text{by definition of } \varepsilon_{1,2} \\ \text{iff } & w_1 \in [\Delta_{\varepsilon_{1,2}}(\sigma(\vec{\varphi}))] && \text{by } \mathrm{M}_{\mathsf{ANON}} \\ \text{iff } & w_1 \leq_{\sigma(\vec{\varphi})} w_2, \end{aligned}$$

and hence $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{ANON}}$. Conversely, if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{ANON}}$, then, for any profile $\vec{\varphi}$, propositional formula μ and permutation σ, it holds that:

$$\begin{aligned} [\Delta_\mu(\vec{\varphi})] &= \min_{\leq_{\vec{\varphi}}}[\mu] && \text{by the fact that } \preccurlyeq \text{ represents } \Delta \\ &= \min_{\leq_{\sigma(\vec{\varphi})}}[\mu] && \text{by } \mathrm{m}_{\mathsf{ANON}} \\ &= [\Delta_\mu(\sigma(\vec{\varphi}))], \end{aligned}$$

and hence Δ satisfies postulate $\mathrm{M}_{\mathsf{ANON}}$.

The proof for Equivalence (2) is essentially similar to the one for Equivalence (1).

Theorem 5.1 shows that postulates $\mathrm{M}_{\mathsf{ANON}}$ and $\mathrm{M}_{\mathsf{NEUT}}$ can be emulated on the semantic level by anonymous and neutral assignments, respectively. But this is only half the battle: from Section 3.4 we know that the standard way of generating assignments for merging operators is to use distances and aggregation functions, so the obvious next question is how to guarantee that assignments generated using these components satisfy properties $\mathrm{m}_{\mathsf{ANON}}$ and $\mathrm{m}_{\mathsf{NEUT}}$. The answer, for $\mathrm{m}_{\mathsf{ANON}}$, turns out to lie with the aggregation function, whereas for $\mathrm{m}_{\mathsf{NEUT}}$ it lies with the distance function.

If $\oplus$ is an aggregation function, then the following property is of interest, for any real numbers $x_1, \ldots, x_n$ and permutation σ of $N = \{1, \ldots, n\}$:

$$(\mathsf{Ag}_{\mathsf{ANON}})\ \ \oplus(x_1, \ldots, x_n) = \oplus(x_{\sigma(1)}, \ldots, x_{\sigma(n)}). \qquad \text{(anonymity)}$$

An aggregation function $\oplus$ is *anonymous* if it satisfies property $\mathsf{Ag}_{\mathsf{ANON}}$. It is now easy to see that if $\oplus$ is anonymous, then any $(d, \oplus)$-induced merging operator $\Delta^{d,\oplus}$ is also anonymous.

Proposition 5.1

If d is a distance between interpretations and $\oplus$ is an aggregation function that satisfies property $\mathsf{Ag}_{\mathsf{ANON}}$, then the $(d, \oplus)$-induced $\mathcal{L}^n$-assignment $\preccurlyeq^{d,\oplus}$ on interpretations satisfies property $\mathrm{m}_{\mathsf{ANON}}$.

Proof

If σ is a permutation of the set $N = \{1, \ldots, n\}$, then we have that:

$$\begin{aligned} w_1 \leq^{d,\oplus}_{\sigma(\vec{\varphi})} w_2 \text{ iff } & d^{\oplus}(\sigma(\vec{\varphi}), w_1) \leq_{\text{lex}} d^{\oplus}(\sigma(\vec{\varphi}), w_2) \\ \text{iff } & \oplus(d(\varphi_{\sigma(i)}, w_1))_{1\leq i\leq n} \leq_{\text{lex}} \oplus(d(\varphi_{\sigma(i)}, w_2))_{1\leq i\leq n} \\ \text{iff } & \oplus(d(\varphi_i, w_1))_{1\leq i\leq n} \leq_{\text{lex}} \oplus(d(\varphi_i, w_2))_{1\leq i\leq n} \qquad \text{by } \mathsf{Ag}_{\mathsf{ANON}} \\ \text{iff } & w_1 \leq^{d,\oplus}_{\vec{\varphi}} w_2. \end{aligned}$$

Proposition 5.1 shows that merging operators induced using anonymous aggregation functions are anonymous, for any distance function. We can even strengthen this result by noticing that anonymity is guaranteed by the standard merging postulates, and in particular postulate M_3, which, we may recall from Section 3.4, says that if two profiles $\vec{\varphi}_1$ and $\vec{\varphi}_2$ are equivalent, in the sense that formulas in $\vec{\varphi}_1$ can be bijectively mapped to equivalent formulas in $\vec{\varphi}_2$, then merging both profiles under equivalent constraints yields equivalent results.

Proposition 5.2

If Δ is an $\mathcal{L}^n$-merging operator that satisfies postulate M_3, then Δ also satisfies postulate $\mathrm{M}_{\mathsf{ANON}}$.

Proof

Note that $\vec{\varphi} \equiv \sigma(\vec{\varphi})$, for any permutation σ. Applying postulate M_3, it immediately follows that $\Delta_\mu(\vec{\varphi}) \equiv \Delta_\mu(\sigma(\vec{\varphi}))$.

Proposition 5.2 only confirms the fact that all the standard merging operators presented in Section 3.4 satisfy the anonymity postulate $\mathrm{M}_{\mathsf{ANON}}$, since, by Proposition 3.8, they all satisfy postulate M_3.

We can approach neutrality in a similar way. Recall, from Section 4.4, that a distance function is *neutral* if, for any interpretations w_1 and w_2 and renaming ρ of atoms, the following property holds:

($\mathrm{D}_{\mathsf{NEUT}}$) $d(w_1, w_2) = d(\rho(w_1), \rho(w_2))$.

Intuitively, property $\mathrm{D}_{\mathsf{NEUT}}$ says that the distance function d is invariant under renamings of atoms. It turns out that if a distance function d is neutral in this sense, then the $(d, \oplus)$-generated assignment satisfies property $\mathrm{m}_{\mathsf{NEUT}}$.

Proposition 5.3

If $\vec{\varphi}$ is a profile, $\oplus$ is an aggregation function and d is a distance function that satisfies property $\mathrm{D}_{\mathsf{NEUT}}$, then the $(d, \oplus)$-generated $\mathcal{L}^n$-assignment $\preccurlyeq^{d,\oplus}$ on interpretations satisfies property $\mathrm{m}_{\mathsf{NEUT}}$.

Proof

If w_1 and w_2 are two interpretations and ρ is a renaming of A, then it holds that:

$$
\begin{aligned}
w_1 \leq_{\vec{\varphi}}^{d,\oplus} w_2 \text{ iff } & d^{\oplus}(\vec{\varphi}, w_1) \leq_{\text{lex}} d^{\oplus}(\vec{\varphi}, w_2) \\
\text{iff } & \oplus(d(\varphi_i, w_1))_{\varphi_i \in \vec{\varphi}} \leq_{\text{lex}} \oplus(d(\varphi_i, w_2))_{\varphi_i \in \vec{\varphi}} \\
\text{iff } & \oplus(\min(d(v, w_1))_{v \in [\varphi_i]})_{\varphi_i \in \vec{\varphi}} \leq_{\text{lex}} \oplus(\min(d(v, w_2))_{v \in [\varphi_i]})_{\varphi_i \in \vec{\varphi}} \\
\text{iff } & \oplus(\min(d(\rho(v), \rho(w_1)))_{\rho(v) \in [\rho(\varphi_i)]})_{\rho(\varphi_i) \in \rho(\vec{\varphi})} \leq_{\text{lex}} \\
& \qquad \oplus(\min(d(\rho(v), \rho(w_2)))_{\rho(v) \in [\rho(\varphi_i)]})_{\rho(\varphi_i) \in \rho(\vec{\varphi})} \\
\text{iff } & \oplus(d(\rho(\varphi_i), \rho(w_1)))_{\rho(\varphi_i) \in \rho(\vec{\varphi})} \leq_{\text{lex}} \oplus(d(\rho(\varphi_i), \rho(w_2)))_{\rho(\varphi_i) \in \rho(\vec{\varphi})} \\
\text{iff } & d^{\oplus}(\rho(\vec{\varphi}), \rho(w_1)) \leq_{\text{lex}} d^{\oplus}(\rho(\vec{\varphi}), \rho(w_2)) \\
\text{iff } & \rho(w_1) \leq_{\rho(\vec{\varphi})}^{d,\oplus} \rho(w_2).
\end{aligned}
$$

Thus, $\leq_{\vec{\varphi}}^{d,\oplus}$ satisfies property for any profile $\vec{\varphi}$ and renaming ρ.

It is straightforward to see that the Hamming distance d_{H} and the drastic distance d_{D} both satisfy property $\text{D}_{\textsf{NEUT}}$. This, together with Propositions 5.3 and 5.2 and a counterexample for postulate $\text{M}_{\textsf{BNDL}}$, completes the picture for the merging operators we are interested in.

Proposition 5.4

The merging operators $\Delta^{\text{H},\oplus}$ and $\Delta^{\text{D},\oplus}$, for $\oplus \in \{\text{sum}, \text{leximax}, \text{leximin}\}$, all satisfy postulates $\text{M}_{\textsf{ANON}}$ and $\text{M}_{\textsf{NEUT}}$. Neither of these merging operators satisfies postulate $\text{M}_{\textsf{BNDL}}$.

Proof

For postulate $\text{M}_{\textsf{ANON}}$, use Proposition 5.2 and the fact that the merging operators under consideration satisfy postulates M_{0-8} (and, a fortiori, postulate M_3).

For postulate $\text{M}_{\textsf{NEUT}}$, the conclusion is implied by Propositions 5.3 and the fact that the Hamming distance d_{H} and the drastic distance d_{D} both satisfy property $\text{D}_{\textsf{NEUT}}$.

For postulate $\text{M}_{\textsf{BNDL}}$, take $\varphi_1 = a$, $\varphi_2 = \neg b$, and $\mu = \top$. Compare $\Delta^{d,\oplus}_{\mu \wedge (a \leftrightarrow b)}(\varphi_1, \varphi_2)$ with $\Delta^{d,\oplus}_{\mu^{a \rightsquigarrow b}}(\varphi_1^{a \rightsquigarrow b}, \varphi_2^{a \rightsquigarrow b})$. This works for both distances and all aggregation functions.

5.2 Collective efficiency

The postulates in this section try to make sure that the merging process cannot be highjacked by a group of agents that, conceivably, are not representative of the whole profile. We attempt to adapt popular properties used to understand voting rules, such as Pareto efficiency, non-dictatorship and the notions of a majority, or Condorcet winner: properties that, to some extent, implement a notion of efficiency at the social level.

As expected, some preliminary notions need to be introduced before the main notions can become intelligible, with none more demanding than the notions surrounding the Condorcet winner. In a voting scenario, finding Condorcet winners involves querying voters on their preference over two candidates at a time, i.e., reducing the election to a contest between two candidates. Adapting this to the context of merging, where candidates are interpretations, would mean finding out how an agent ranks any two outcomes. Since the input to a merging operator is a profile of formulas (i.e., sets of interpretations) and not rankings over alternatives, this might seem like an unnatural suggestion. However, since we are working with merging operators that satisfy postulates M_{0-8}, Theorem 3.10 shows that we can always narrow down the question to just two interpretations, using an $\mathcal{L}$-proxy. Recall from Section 2.1 that if w_1 and w_2 are two interpretations, an $\mathcal{L}$-proxy of $\{w_1, w_2\}$ is a propositional formula $\varepsilon_{1,2}$ such that $[\varepsilon_{1,2}] = \{w_1, w_2\}$. Taking, then, $\varepsilon_{1,2}$ as the constraint for a merging task effectively gives us the ranking of w_1 and w_2 relative to a profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$, using the Δ-revealed relation. Recall, from Section 3.4 that, given a merging operator Δ and interpretations w_1 and w_2, the Δ-revealed ranking $\leq^{\Delta}_{\varphi}$ is defined as:

$$w_1 \leq^{\Delta}_{\vec{\varphi}} w_2 \text{ if } w_1 \in [\Delta_{\varepsilon_{1,2}}(\vec{\varphi})].$$

Intuitively, this means that according to the profile $\vec{\varphi}$, outcome w_1 is considered at least as good as outcome w_2. Furthermore, taking the profile to consist of only one formula φ_i, i.e., $\vec{\varphi} = (\varphi_i)$, gives us the preference of the agent i over w_1 and w_2. Thus, as previously established, if the profile is (φ_i), we write $\leq_{\varphi_i}$ instead of $\leq_{(\varphi_i)}$. With this wisdom in hand, we can now go ahead and adapt the notion of a Condorcet winner to the context of merging.

If $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ is an $\mathcal{L}$-profile, μ is a propositional formula, Δ is an $\mathcal{L}^n$-merging operator, and w_1 and w_2 are models of μ, the *support* $\mathrm{supp}_{\mu}(w_1, w_2)$ *of* w_1 *over* w_2 *with respect to* $\vec{\varphi}$ *and* μ is defined as:

$$\mathrm{supp}_{\mu}(w_1, w_2) = \{i \in N \mid w_1 \leq^{\Delta}_{\varphi_i} w_2\},$$

i.e., the set of agents in N whose point of view φ_i implies, through the Δ-revealed relation, that w_1 is at least as good as w_2 according to $\leq^{\Delta}_{\varphi_i}$. If the merging operator Δ is generated using a distance function d and an aggregation function $\oplus$, we write simply $\mathrm{supp}^{d,\oplus}_{\mu}(w_1, w_2)$. The *size of the support of* w_1 *over* w_2 *with respect to* $\vec{\varphi}$ *and* μ is $|\mathrm{supp}_{\mu}(w_1, w_2)|$, i.e., the number of agents who think w_1 is at least as good as w_2. If $w^* \in [\mu]$, then w^* is a *weak Condorcet winner with respect to* $\vec{\varphi}$ *and* μ if it holds that:

$$|\mathrm{supp}_{\mu}(w^*, w)| \geq |\mathrm{supp}_{\mu}(w, w^*)|, \text{ for any } w \in [\mu],$$

i.e., if the size of the support for w^* over any other model w of μ is at least as large as the size of the support of w over w^* (with respect to $\vec{\varphi}$ and μ). In a break with standard practice, we write $\mathsf{COND}_\mu(\vec{\varphi})$ for an $\mathcal{L}$-proxy of the set of weak Condorcet winners with respect to $\vec{\varphi}$ and μ, i.e., is $\mathsf{COND}_\mu(\vec{\varphi})$ is a propositional formula such that: $\mathsf{COND}_\mu(\vec{\varphi}) = \{w^* \in [\mu] \mid w^* \text{ is a weak Condorcet winner with respect to } \vec{\varphi} \text{ and } \mu\}$.

We can use the same tactic to define an even stronger notion, which requires an interpretation to be in the top choices of at least half of all agents, relative to a constraint μ. If $\vec{\varphi}$ is a propositional profile, w is an interpretation and μ is a propositional formula, the *support* $\mathrm{supp}(w, \mu)$ *of* w *over* μ *with respect to* $\vec{\varphi}$ is defined as:

$$\mathrm{supp}(w, \mu) = \{i \in N \mid w \leq^{\Delta}_{\varphi_i} w', \text{ for every } w' \in [\mu]\},$$

i.e., the set of agents in N for whom w is at least as good as all the models of μ, according to the Δ-revealed ranking $\leq^{\Delta}_{\varphi_i}$. Note that if w is itself a model of μ, then $\mathrm{supp}(w, \mu)$ gathers all the agents for whom $w \in \min_{\leq^{\Delta}_{\varphi_i}}[\mu]$, i.e., for whom w is a top choice when the menu is restricted to the models of μ. We write simply $\mathrm{supp}^{d,\oplus}(w, \mu)$ if the merging operator Δ is generated using a distance function d and an aggregation function $\oplus$. If μ is a constraint and $w \in [\mu]$, then w is *majority-supported with respect to* $\vec{\varphi}$ *and* μ if it holds that:

$$|\mathrm{supp}(w, \mu)| \geq \lfloor \frac{n}{2} \rfloor + 1,$$

i.e., if the size of the support for w over μ with respect to $\vec{\varphi}$ is at least half of all the agents in $\vec{\varphi}$. Intuitively, w is majority-supported with respect to $\vec{\varphi}$ and μ if w is a model of μ and a majority of the agents in $\vec{\varphi}$ find at least as good as any other model of μ. Consequently, the condition of w being majority-supported with respect to $\vec{\varphi}$ and μ can be rewritten as saying that $w \in [\Delta_\mu(\varphi_i)]$, for a majority of the formulas φ_i in $\vec{\varphi}$. In another break with standard practice, we will denote by $\mathsf{MAJR}_\mu(\vec{\varphi})$ an $\mathcal{L}$-proxy for the set of majority-supported outcomes with respect to $\vec{\varphi}$ and μ, i.e., $\mathsf{MAJR}_\mu(\vec{\varphi})$ is a propositional formula such that $[\mathsf{MAJR}_\mu(\vec{\varphi})] = \{w \in [\mu] \mid w \text{ is majority-supported with respect to } \vec{\varphi} \text{ and } \mu\}$.

We can now introduce the actual postulates, which, unless otherwise stated, are intended to apply for any set of agents $N = \{1, \ldots, n\}$, profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ and constraints μ, μ_1 and μ_2:

($\mathrm{M}_{\mathsf{NOND}}$) There is no agent i in N such that, for any profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ and constraint μ, it holds that $\Delta_\mu(\vec{\varphi}) \equiv \Delta_\mu(\varphi_i)$.

($\mathrm{M}_{\mathsf{wPAR}}$) $\Delta_\mu(\varphi_1) \wedge \cdots \wedge \Delta_\mu(\varphi_n) \models \Delta_\mu(\vec{\varphi})$.

($\mathrm{M}_{\mathsf{sPAR}}$) If $\Delta_\mu(\varphi_1) \wedge \cdots \wedge \Delta_\mu(\varphi_n)$ is consistent, then $\Delta_\mu(\vec{\varphi}) \models \Delta_\mu(\varphi_1) \wedge \cdots \wedge \Delta_\mu(\varphi_n)$.

($\mathrm{M}_{\mathsf{CSOV}}$) For any propositional formula μ_2, there exists a profile $\vec{\varphi}$ such that $\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2 \equiv \mu_1 \wedge \mu_2$.

($\mathrm{M}_{\mathsf{COND}}$) $\mathsf{COND}_\mu(\vec{\varphi}) \models \Delta_\mu(\vec{\varphi})$.

($\mathrm{M}_{\mathsf{MAJR}}$) $\mathsf{MAJR}_\mu(\vec{\varphi}) \models \Delta_\mu(\vec{\varphi})$.

Postulate $\mathrm{M}_{\mathsf{NOND}}$, where '$\mathsf{NOND}$' stands for *non-dictatorship*, prevents the extreme case where one agent has the power to skew the result in its direction. This property is analogous to the eponymous property in voting, where it is satisfied if there is no single voter who, alone, is able to determine the outcome of an election. In a social choice scenario non-dictatorhip is usually featured as a minimal requirement that any reasonable voting method should satisfy, and we view it in similar terms in merging.

On the other end of the spectrum, postulates $\mathrm{M}_{\mathsf{wPAR}}$ and $\mathrm{M}_{\mathsf{sPAR}}$, where '$\mathsf{wPAR}$' and '$\mathsf{sPAR}$' stand for *weak* and *strong Pareto*, respectively, address the limit case where there is anonymous agreement over certain outcomes, and they ensure that this agreement is reflected in the result. The weak Pareto postulate $\mathrm{M}_{\mathsf{wPAR}}$ says that any outcomes universally agreed upon should be part of the result, while the strong Pareto postulate $\mathrm{M}_{\mathsf{sPAR}}$ says that *only* unversally agreed upon outcomes are part of the result, if such outcomes exist. These postulates have made an appeareance before in the merging literature: postulates M_{0-4} and M_{7-8} together with $\mathrm{M}_{\mathsf{wPAR}}$ and $\mathrm{M}_{\mathsf{sPAR}}$ characterize what is called a *pre-IC merging operator* [Everaere et al., 2014], and it has already been noted that any merging operator satisfying postulates M_{0-8} is also a pre-IC merging operator, as defined just now.

Postulate $\mathrm{M}_{\mathsf{CSOV}}$, where '$\mathsf{CSOV}$' stands for *citizen sovereignty*, is modeled after an eponymous voting property requiring that for any candidate c there is at least one profile of votes according to which c is the winner, i.e., no candidate is *a priori* denied a seat at the winning table. Since the merging task is parameterized by a constraint μ_1 that has to be satisfied, the citizen sovereignty becomes the requirement that any selection of models of μ_1 should be within the reach of a merging operator: this selection is realized by a propositional formula μ_2, which is added to μ_1 to single out the models of μ_1 of interest. thus, postulate $\mathrm{M}_{\mathsf{CSOV}}$ can be read as saying that any outcomes consistent with μ_1 that also happen to be models of μ_2 can form the result when μ_1 is the constraint.

The notion of a weak Condorcet winner is a straight adaptation of the eponymous notion formulated in Section 2.4 for social choice functions: if the support of an outcome w over w' with respect to $\vec{\varphi}$ and μ is at least as large as the support of w' over w, then this is like w winning over (or tying with) w' in a head-to-head election. If w manages to win over any other outcome in $[\mu]$ (or, at least not to lose), then this is taken as a strong reason to include w in the result: the notion is 'weak' because it admits ties. Postulate $\mathrm{M}_{\mathsf{COND}}$, then, where '$\mathsf{COND}$' stands for *Condorcet*, says that the result should include any outcomes that are weak Condorcet winners. Note that the postulate is trivially satisfied if the set of weak Condorcet winners is empty.

The notion of a majority-supported outcome w with respect to a profile $\vec{\varphi}$ and a constraint μ is intended to capture outcomes that are the most preferred outcomes of a majority of the agents. Note that, since the merging operator Δ is assumed to satisfy postulates M_{0-8}, the Δ-induced ranking $\leq^{\Delta}_{\vec{\varphi}}$ is, by Theorem 3.10, a total preorder on interpretations.

Revisiting the definition of a majority-supported outcome w with respect to $\vec{\varphi}$ and μ, we can see that w being $\leq^{\Delta}_{\varphi_i}$-preferred to every model of μ is the same as saying that $w \in \min_{\leq^{\Delta}_{\varphi_i}}[\mu]$, i.e., that $w \in [\Delta_\mu(\varphi_i)]$. In other words, if we see the merging process as an election over the models of μ, then a majority-supported outcome w is the top choice in a majority of the preferences. Postulate $\mathrm{M}_{\mathsf{MAJR}}$, where 'MAJR' stands for *majority*, then says that w, along with all the other outcomes that share this property, have to be among the winning outcomes. Note that postulate $\mathrm{M}_{\mathsf{MAJR}}$ is different from postulate $\mathrm{M}_{\mathsf{MAJ}}$ in Section 3.4.

Example 5.4: Majority supported outcomes and weak Condorcet winners

For the merging scenario in Example 5.1 we have that $[\mu] = \{ab, ac, bc\}$, hence we need only look at how a merging operator ranks these three outcomes. For the merging operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$ we obtain that $\mathrm{supp}^{\mathrm{H},\,\mathrm{sum}}_{\mu}(ab, ac) = \{1, 2, 3\}$, $\mathrm{supp}^{\mathrm{H},\,\mathrm{sum}}_{\mu}(ac, ab) = \{4\}$, $\mathrm{supp}^{\mathrm{H},\,\mathrm{sum}}_{\mu}(ab, bc) = \{1, 2, 3\}$, $\mathrm{supp}^{\mathrm{H},\,\mathrm{sum}}_{\mu}(bc, ab) = \{3, 4\}$, $\mathrm{supp}^{\mathrm{H},\,\mathrm{sum}}_{\mu}(bc, ac) = \{1, 3, 4\}$ and $\mathrm{supp}^{\mathrm{H},\,\mathrm{sum}}_{\mu}(ac, bc) = \{1, 2, 3\}$. Consequently, ab is the only majority-supported outcome with respect to $\vec{\varphi}$ and μ, since it is the only outcome present in the top choices of three or more agents; ab is also the only weak Condorcet winner, as no other outcome in $[\mu]$ beats it in a head-to-head contest of support.

As in the more traditional social choice settings, weak Condorcet winners and majority-supported outcomes are not guaranteed to exist, particularly if majority cycles are present. For merging operators generated using distances, the existence of such cycles depends on whether the distance functions manage to arrange things just right so as to induce the right kind of preference order in a profile of distance-based preferences. The following example shows that this is eminently possible when the distance used is the Hamming distance.

Example 5.5: Weak Condorcet winners do not always exist

For the set of atoms $A = \{a, b, c, d\}$, consider a profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq 3}$, with $[\varphi_1] = \{a\}$, $[\varphi_2] = \{acd\}$ and $[\varphi_3] = \{bc\}$, and a constraint μ, with $[\mu] = \{a, b, cd\}$. The Hamming distances from each of the formulas in $\vec{\varphi}$ to each of the models of μ, together with the preorders $\leq^{\mathrm{H},\,\oplus}_{\varphi_i}$ on the models of μ, for $i \in \{1, 2, 3\}$, are depicted in Figure 5.1.

Note that there is no weak Condorcet winner with respect to $\vec{\varphi}$ and μ. For instance, the support of a over b is $\mathrm{supp}^{\mathrm{H},\,\oplus}_{\mu}(a, b) = \{1, 2\}$, while the support of b over a is $\mathrm{supp}^{\mathrm{H},\,\oplus}_{\mu}(b, a) = \{3\}$, which means that a beats b in a head to head election. However, $\mathrm{supp}^{\mathrm{H},\,\oplus}_{\mu}(a, cd) = \{1\}$ and $\mathrm{supp}^{\mathrm{H},\,\oplus}_{\mu}(cd, a) = \{2, 3\}$, which means that a loses to cd in a head to head contest. The same holds for all other pairs of models of μ. Likewise, there is no majority-supported outcome with respect to $\vec{\varphi}$ and μ, since none of the models of μ is in the top choices of more than two of the agents in $\vec{\varphi}$.

Example 5.5 shows that the postulates just introduced are best understood by looking

d_{H}	$[\varphi_1]$ $\{a\}$	$[\varphi_2]$ $\{acd\}$	$[\varphi_3]$ $\{bc\}$
a	0	2	3
b	2	4	1
cd	3	1	2

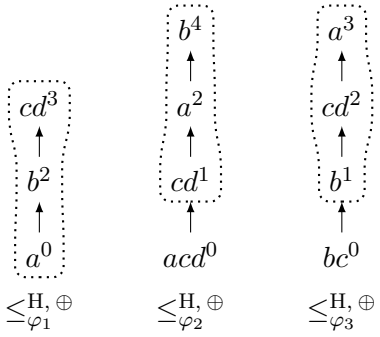


Figure 5.1: Preorders generated using Hamming distances and an aggregation function $\oplus$ for the profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq 3}$. A majority cycle between a, b and cd means that there is no weak Condorcet winner with respect to $\vec{\varphi}$ and μ, for $[\mu] = \{a, b, cd\}$. The fact that none of the models of μ is in the top choices of more than two agents in $\vec{\varphi}$ means that there is no majority-supported outcome with respect to $\vec{\varphi}$ and μ outcome either.

at what they expect of the preorders describing the profile. To make this connection more precise, we present a set of properties meant to apply to an $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations that represents an $\mathcal{L}^n$-merging operator. To make sense of the following properties, recall from Section 2.4 that a weak Condorcet winner with respect to a preference profile and a set of alternatives is an alternative that gets at least as much support as every other alternative in the set. The following properties are meant to apply for any set $N = \{1, \ldots, n\}$ of agents, $\mathcal{L}$-profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$, interpretations w_1 and w_2 and sets of interpretations $\mathcal{W}$, $\mathcal{W}_1$, $\mathcal{W}_2$:

($\mathrm{m}_{\mathsf{NOND}}$) There is no agent i such that $\leq_{\vec{\varphi}} = \leq_{\varphi_i}$, for any profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$.

($\mathrm{m}_{\mathsf{wPAR}}$) If $w_1 \leq_{\varphi_i} w_2$, for all $i \in N$, then $w_1 \leq_{\vec{\varphi}} w_2$.

($\mathrm{m}_{\mathsf{sPAR}}$) If $w_1 \leq_{\varphi_i} w_2$, for all $i \in N$, and there exists $j \in N$ such that $w_1 <_{\varphi_j} w_2$, then $w_1 <_{\vec{\varphi}} w_2$.

($\mathrm{m}_{\mathsf{CSOV}}$) There exists a profile $\vec{\varphi}$ such that $w_1 \leq_{\vec{\varphi}} w_2$, for any $w_1 \in \mathcal{W}_1$ and $w_2 \in \mathcal{W}_2$.

($\mathrm{m}_{\mathsf{COND}}$) If w is a weak Condorcet winner with respect to $(\leq_{\varphi_i})_{1\leq i\leq n}$ and $\mathcal{W}$, then $w \leq_{\vec{\varphi}} w'$, for any $w' \in \mathcal{W}$.

($\mathrm{m}_{\mathsf{MAJR}}$) If $w_1 \leq_{\varphi_i} w_2$ for a majority of $i \in N$, then $w_1 \leq_{\vec{\varphi}} w_2$.

An $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations is *non-dictatorial*, *weak* and *strong* Pareto efficient, *Condorcet consistent* and *majority consistent* if it satisfies properties $\mathrm{m}_{\mathsf{NOND}}$, $\mathrm{m}_{\mathsf{wPAR}}$, $\mathrm{m}_{\mathsf{sPAR}}$, $\mathrm{m}_{\mathsf{COND}}$ and $\mathrm{m}_{\mathsf{MAJR}}$, respectively. The properties just introduced map neatly onto the postulates presented earlier. Since these postulates were tailored specifically to capture solution concepts from voting theory, this comes as no surprise.

Theorem 5.2

If Δ is an $\mathcal{L}^n$-merging operator that satisfies postulates M_{0-1} and M_3 and $\preccurlyeq$ is a total $\mathcal{L}^n$-assignment on interpretations that represents it, then the following equivalences hold:

(1) Δ satisfies postulate $\mathrm{M}_{\mathsf{NOND}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{NOND}}$.

(2) Δ satisfies postulate $\mathrm{M}_{\mathsf{wPAR}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{wPAR}}$.

(3) Δ satisfies postulate $\mathrm{M}_{\mathsf{sPAR}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{sPAR}}$.

(4) Δ satisfies postulate $\mathrm{M}_{\mathsf{CSOV}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{CSOV}}$.

(5) Δ satisfies postulate $\mathrm{M}_{\mathsf{COND}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{COND}}$.

(6) Δ satisfies postulate $\mathrm{M}_{\mathsf{MAJR}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{MAJR}}$.

Proof

For a comment on the role of postulates M_{0-1} and M_3, see the comment at the beginning of the proof for Theorem 5.1.

For Equivalence (1), we have that the existence of an agent i such that $\Delta_\mu(\vec{\varphi}) = \Delta_\mu(\varphi_i)$, for any profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$ and constraint μ, is equivalent to the fact that $w_1 \leq_{\vec{\varphi}} w_2$ if and only if $w_1 \leq_{\varphi_i} w_2$, for any interpretations w_1 and w_2, i.e., to the fact that $\leq_{\vec{\varphi}} = \leq_{\varphi_i}$.

For Equivalence (2), suppose first that Δ satisfies postulate $\mathrm{M}_{\mathsf{wPAR}}$ and assume that $\preccurlyeq$ does not satisfy property $\mathrm{m}_{\mathsf{wPAR}}$. This means that there exist interpretations w_1 and w_2 such that $w_1 \leq_{\varphi_i}$, for $i \in N$, and $w_2 \leq_{\vec{\varphi}} w_1$. Taking the constraint $\varepsilon_{1,2}$ yields a contradiction with postulate $\mathrm{M}_{\mathsf{wPAR}}$. Conversely, suppose $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{wPAR}}$ and Δ does not satisfy postulate $\mathrm{M}_{\mathsf{wPAR}}$. This implies that there exists $w_1 \in [\Delta_\mu(\varphi_1) \wedge \cdots \wedge \Delta_\mu(\varphi_n)]$ such that $w_1 \notin [\Delta_\mu(\vec{\varphi})]$. The latter conclusion implies, by postulates M_0 and M_1 and the assumption that $\leq_{\vec{\varphi}}$ is total, that there exists an interpretation $w_2 \in [\Delta_\mu(\vec{\varphi})]$ such that $w_2 <_{\vec{\varphi}} w_1$. The former conclusion, however, implies that $w_1 \leq_{\varphi_i} w_2$, for any $i \in N$, which, together with property $\mathrm{m}_{\mathsf{wPAR}}$, implies that $w_1 \leq_{\vec{\varphi}} w_2$. We have thus arrived at a contradiction.

For Equivalence (3), suppose first that Δ satisfies postulate $\mathrm{M}_{\mathsf{sPAR}}$ and assume that $\preccurlyeq$ does not satisfy property $\mathrm{m}_{\mathsf{sPAR}}$. This means that there exist interpretations w_1 and w_2 such that $w_1 \leq_{\varphi_i} w_2$, for $i \in N$, $w_1 <_j w_2$, for some $j \in N$ and, furthermore, that $w_2 \leq_{\vec{\varphi}} w_1$. This means that $\Delta_{\varepsilon_{1,2}}(\varphi_1) \wedge \cdots \wedge \Delta_{\varepsilon_{1,2}}(\varphi_n)$ is consistent, which, by postulate $\mathrm{M}_{\mathsf{sPAR}}$, implies that $w_2 \in [\Delta_{\varepsilon_{1,2}}(\varphi_j)]$. But this is a contradiction. Conversely, suppose $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{sPAR}}$ and Δ does not satisfy postulate $\mathrm{M}_{\mathsf{sPAR}}$. This implies that there exists an interpretation $w_1 \in [\Delta_\mu(\vec{\varphi})]$ such that

$w_1 \notin [\Delta_\mu(\varphi_1) \wedge \cdots \wedge \Delta_\mu(\varphi_n)]$. The latter conclusion, together with postulates M_0, M_1 and the assumption that $\preccurlyeq$ is total, implies that there exists $j \in N$ such that $w_1 \notin [\Delta_\mu(\varphi_j)]$. The assumption that $\Delta_\mu(\varphi_1) \wedge \cdots \wedge \Delta_\mu(\varphi_n)$ is consistent implies that there exists $w_2 \in [\Delta_\mu(\varphi_1) \wedge \cdots \wedge \Delta_\mu(\varphi_n)]$. Putting the last two facts together implies that $w_2 \leq_{\varphi_i} w_1$, for every $i \in N$, and $w_2 <_{\varphi_j} w_1$, which, by property m_{sPAR}, yields that $w_2 <_{\vec{\varphi}} w_1$. But this contradicts the fact that $w_1 \in [\Delta_\mu(\vec{\varphi})]$.

For Equivalence (4), the statement is trivially true if $\mu_1 \wedge \mu_2$ is inconsistent, hence we only look at the case when $\mu_1 \wedge \mu_2$ is consistent. For one direction, suppose Δ satisfies postulate M_{CSOV}: then, for any sets $\mathcal{W}_1$ and $\mathcal{W}_2$ of interpretations, take the $\mathcal{L}$-proxies of $\mathcal{W}_1 \cup \mathcal{W}_2$ and $\mathcal{W}_1$, i.e., two propositional formulas $\varepsilon_{\mathcal{W}_1 \cup \mathcal{W}_2}$ and $\varepsilon_{\mathcal{W}_1}$ such that $[\varepsilon_{\mathcal{W}_1 \cup \mathcal{W}_2}] = \mathcal{W}_1 \cup \mathcal{W}_2$ and $[\varepsilon_{\mathcal{W}_1}] = \mathcal{W}_1$. Using postulate M_{CSOV}, we have that there exists a profile $\vec{\varphi}$ such that $\Delta_{\varepsilon_{\mathcal{W}_1 \cup \mathcal{W}_2}}(\vec{\varphi}) \wedge \varepsilon_{\mathcal{W}_1} \equiv \varepsilon_{\mathcal{W}_1 \cup \mathcal{W}_2} \wedge \varepsilon_{\mathcal{W}_1}$, which implies that $[\Delta_{\varepsilon_{\mathcal{W}_1 \cup \mathcal{W}_2}}(\vec{\varphi})] = \mathcal{W}_1$. This, in turn, implies that $\min_{\leq_{\vec{\varphi}}}[\varepsilon_{\mathcal{W}_1 \cup \mathcal{W}_2}] = \mathcal{W}_1$, from which the conclusion follows. Conversely, we take $\mathcal{W}_1 = [\mu_1 \wedge \mu_2]$ and $\mathcal{W}_2 = [\mu_1]$.

For Equivalence (5), we remark that $w_1 \leq^{\Delta}_{\vec{\varphi}} w_2$ is equivalent to $w_1 \in [\Delta_{\varepsilon_{1,2}}(\vec{\varphi})]$, which is equivalent to the fact that $w_1 \in \min_{\leq_{\vec{\varphi}}}[\varepsilon_{1,2}]$, or $w_1 \leq_{\vec{\varphi}} w_2$, for any total preorder $\leq_{\vec{\varphi}}$ used to represent Δ. This implies that an interpretation w^* being a weak Condorcet winner with respect to $(\leq_{\varphi_i})_{1 \leq i \leq n}$ and μ is equivalent to w^* being a weak Condorcet winner with respect to the preference profile $\vec{\leq} = (\leq_{\varphi_i})_{1 \leq i \leq n}$ and $[\mu]$. If Δ is an $\mathcal{L}^n$-merging operator that satisfies postulate M_{COND}, then a weak Condorcet winner w^* with respect to $\vec{\leq}$ and a set of interpretations $\mathcal{W}$ will be a model of $\Delta_{\varepsilon_{\mathcal{W}}}(\vec{\varphi})$, and this implies that $w^* \in \min_{\leq_{\vec{\varphi}}}[\varepsilon_{\mathcal{W}}]$, i.e., that $w^* \leq_{\vec{\varphi}} w$, for any interpretation $w \in [\mathcal{W}]$. Conversely, a weak Condorcet winner w^* with respect to $\vec{\varphi}$ and μ is a weak Condorcet winner with respect to $(\leq_{\varphi_i})_{1 \leq i \leq n}$ and $[\mu]$ and, by property m_{COND}, it holds that $w^* \leq_{\vec{\varphi}} w$, for any $w \in [\mu]$, which implies that $w^* \in \min_{\leq_{\vec{\varphi}}}[\mu]$, or $w^* \in [\Delta_\mu(\vec{\varphi})]$.

For Equivalence (6) we use the same observation as above to argue that w^* being a majority-supported outcome with respect to $\vec{\varphi}$ and μ is equivalent to w^* being a majority-supported outcome with respect to $(\leq_{\varphi_i})_{1 \leq i \leq n}$ and $[\mu]$, which then yields the conclusion.

Theorem 5.2 makes it clear what an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations needs to look like if a merging operator Δ represented by it is to satisfy the postulates introduced in this section. The immediate next question, however, is whether existing distance-based merging operators actually manage to select majority-supported outcomes, when they exist, or weak Condorcet winner, when they exist, or whether they are non-dictatorial or resolvable. For some of these properties there exists a useful shortcut, since it turns out that they follow directly from postulate M_{0-8}.

Proposition 5.5

If Δ is an $\mathcal{L}^n$-merging operator that satisfies postulates M_{0-8}, then Δ also satisfies postulates $\mathrm{M}_{\mathsf{NOND}}$, $\mathrm{M}_{\mathsf{CSOV}}$, $\mathrm{M}_{\mathsf{wPAR}}$ and $\mathrm{M}_{\mathsf{sPAR}}$.

Proof

For postulate $\mathrm{M}_{\mathsf{NOND}}$, suppose agent 1, with beliefs φ_1, is a dictator for the merging operator Δ. Choose a formula φ_2 such that $\varphi_1 \wedge \varphi_2$ is inconsistent and a constraint $\mu = \varphi_1 \vee \varphi_2$. Since agent 1 is a dictator, we have that $\Delta_\mu(\varphi_1, \varphi_2) \equiv \Delta_\mu(\varphi_1)$. Since $\mu \wedge \varphi_1$ is consistent, by postulate M_2 it follows that $\Delta_\mu(\varphi_1) \equiv \varphi_1 \wedge \mu \equiv \varphi_1$. At the same time we have that $\Delta_\mu(\varphi_1, \varphi_2) \wedge \varphi_1$ is consistent, and thus, by postulate M_4, it holds that $\Delta_\mu(\varphi_1, \varphi_2) \wedge \varphi_2$ is consistent as well. We then have a contradiction with the fact that $\varphi_1 \wedge \varphi_2$ is inconsistent.

For postulate $\mathrm{M}_{\mathsf{CSOV}}$, if $\mu_1 \wedge \mu_2$ is inconsistent, the conclusion is immediate. If $\mu_1 \wedge \mu_2$ is consistent, take a profile $\vec{\varphi} = (\varphi)$, where $\varphi = \mu_1 \wedge \mu_2$. Clearly, $\varphi \wedge \mu_1$ is consistent, hence by postulate M_2 it follows that $\Delta_\mu(\vec{\varphi}) \equiv \varphi \wedge \mu_1 \equiv \mu_1 \wedge \mu_2$.

Postulates $\mathrm{M}_{\mathsf{wPAR}}$ and $\mathrm{M}_{\mathsf{sPAR}}$ follow directly from postulates M_5 and M_6.

In the case of postulate $\mathrm{M}_{\mathsf{MAJR}}$, the situation turns out to be different: the presence of postulate M_2 actually precludes any merging operator from satisfying $\mathrm{M}_{\mathsf{MAJR}}$.

Proposition 5.6

If Δ is an $\mathcal{L}^n$-merging operator that satisfies postulate M_2, then Δ does not satisfy postulate $\mathrm{M}_{\mathsf{MAJR}}$.

Proof

Take a profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq 3}$, with $[\varphi_1] = [\varphi_2] = \{\emptyset, a\}$ and $[\varphi_3] = \{\emptyset\}$, and a constraint μ with $[\mu] = \{\emptyset, a\}$. Postulate M_2 implies that $[\Delta_\mu(\vec{\varphi})] = [\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \mu] = \{\emptyset\}$. Thus, even though a is a top choice of two out of the three agents, a does not make the list of winning interpretations.

The idea behind Proposition 5.6 is that, under postulate M_2, any agent has veto power over an interpretation w: by not including w in its top choices, i.e., by not making w a model of its submitted opinion, the agent makes sure that w is not part of the result: and this will happen even if w is supported by a majority of the agents. Incidentally, postulate M_2 precludes the possibility that anything along the lines of a *plurality-supported outcome* will be guaranteed to be in the result.

We now have all the pieces of information we need to determine where the main distance-

d_{H}	$[\varphi_{1-3}]$ $3\cdot\{a\}$	$[\varphi_{4-5}]$ $2\cdot\{bc\}$	$[\varphi_6]$ $\{b\}$	$[\varphi_7]$ $\{acd\}$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi},\bullet)$	$d_{\mathrm{H}}^{\mathrm{leximax}}(\vec{\varphi},\bullet)$
a	$3\cdot 0$	$2\cdot 3$	2	2	10	$(3,3,2,2,0,0,0)$
b	$3\cdot 1$	$2\cdot 1$	0	4	**9**	$(4,1,1,1,1,1,0)$
cd	$3\cdot 2$	$2\cdot 2$	3	1	14	$(\mathbf{3},\mathbf{2},\mathbf{2},\mathbf{2},\mathbf{2},\mathbf{2},\mathbf{1})$

Table 5.1: Outcome a is the only weak Condorcet winner with respect to the profile $\vec{\varphi}=(\varphi_i)_{1\leq i\leq 7}$ and μ, where $[\mu]=\{a,b,cd\}$, but a is selected by neither $\Delta^{\mathrm{H},\,\mathrm{sum}}$ nor by $\Delta^{\mathrm{H},\,\mathrm{leximax}}$.

based operators stand in relation to the postulates introduced in this section.

Proposition 5.7

If $\oplus$ is either the sum, leximax or leximin aggregation function, then the following statements hold:

(1) postulates $\mathrm{M}_{\mathsf{NOND}}$, $\mathrm{M}_{\mathsf{wPAR}}$, $\mathrm{R}_{\mathsf{sPAR}}$ and $\mathrm{M}_{\mathsf{CSOV}}$ are satisfied by all operators $\Delta^{\mathrm{H},\oplus}$ and $\Delta^{\mathrm{D},\oplus}$;

(2) postulate $\mathrm{M}_{\mathsf{COND}}$ is satisfied by operators $\Delta^{\mathrm{D},\oplus}$, but by neither of the operators $\Delta^{\mathrm{H},\oplus}$;

(3) postulate $\mathrm{M}_{\mathsf{MAJR}}$ is satisfied by neither of the operators $\Delta^{\mathrm{D},\oplus}$ and $\Delta^{\mathrm{H},\oplus}$.

Proof

Since the operators $\Delta^{\mathrm{H},\oplus}$ and $\Delta^{\mathrm{D},\oplus}$, for $\oplus\in\{\mathrm{sum},\mathrm{leximax},\mathrm{leximin}\}$, satisfy postulates M_{0-8}, then, by Proposition 5.5, they also satisfy postulates $\mathrm{M}_{\mathsf{NOND}}$, $\mathrm{M}_{\mathsf{wPAR}}$, $\mathrm{R}_{\mathsf{sPAR}}$, $\mathrm{M}_{\mathsf{CSOV}}$, and by Proposition 5.6 they do not satisfy postulate $\mathrm{M}_{\mathsf{MAJR}}$. This shows that Statements (1) and (3) hold.

For Statement (2) and operators $\Delta^{\mathrm{D},\oplus}$ recall first that all three operators considered here are equivalent, so proving the claim for $\Delta^{\mathrm{D},\,\mathrm{sum}}$ will suffice. Note, as well, that $d_{\mathrm{D}}^{\mathrm{sum}}(\vec{\varphi},w)$ essentially counts the number of agents in N, who have w as their model, for any interpretation w, and $\Delta^{\mathrm{D},\,\mathrm{sum}}$ selects the interpretations in $[\mu]$ that occur most often as models of agents in N. We have, then, that if w^* is a weak Condorcet winner with respect to φ and μ, then $\mathrm{supp}_{\mu}^{\mathrm{D},\,\mathrm{sum}}(w^*,w)\geq\mathrm{supp}^{\mathrm{D},\,\mathrm{sum}}(w,w^*)$, for any interpretation $w\in[\mu]$. This means that w^* occurs as a model of φ_i for at least as many agents $i\in N$ than any other interpretation $w\in[\mu]$, which, as per the previous observation, implies that $w^*\in[\Delta_{\mu}^{\mathrm{D},\,\mathrm{sum}}(\vec{\varphi})]$.

For Statement (2) and operators $\Delta^{\mathrm{H},\,\mathrm{sum}}$ and $\Delta^{\mathrm{H},\,\mathrm{leximax}}$, take the profile $\vec{\varphi}=(\varphi_i)_{1\leq i\leq 7}$, with $[\varphi_1]=[\varphi_2]=[\varphi_3]=\{a\}$, $[\varphi_4]=[\varphi_5]=\{bc\}$, $[\varphi_6]=\{b\}$, $[\varphi_7]=\{acd\}$,

and a constraint μ such that $[\mu] = \{a, b, cd\}$. The Hamming distances from φ_i, for $1 \leq i \leq 7$, to every model of μ, together with the aggregated distances for the sum and leximax aggregation function, are depicted in Table 5.1.

Note that a is the only weak Condorcet winner with respect to $\vec{\varphi}$ and μ, since the size of its support over b and cd is 4 in both cases. In other words, $[\mathsf{COND}_\mu(\vec{\varphi})] = \{a\}$. However, a is selected by neither $\Delta^{\mathrm{H},\,\mathrm{sum}}$ nor $\Delta^{\mathrm{H},\,\mathrm{leximax}}$, since $[\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi})] = \{b\}$ and $[\Delta_\mu^{\mathrm{H},\,\mathrm{leximax}}(\vec{\varphi})] = \{cd\}$.

For Statement (2) and operator $\Delta^{\mathrm{H},\,\mathrm{leximin}}$, a simpler counterexample will suffice. Take the profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq 3}$, with $[\varphi_1] = [\varphi_2] = \{\emptyset\}$ and $[\varphi_3] = \{ab\}$, and a constraint μ with $[\mu] = \{a, ab\}$. We have that $d_{\mathrm{H}}^{\mathrm{leximin}}(\vec{\varphi}, a) = (1, 1, 1)$ and $d_{\mathrm{H}}^{\mathrm{leximin}}(\vec{\varphi}, ab) = (0, 2, 2)$, which means that $[\Delta_\mu^{\mathrm{H},\,\mathrm{leximin}}(\vec{\varphi})] = \{ab\}$. However, a is the only weak Condorcet winner (and even the majority supported outcome) with respect to $\vec{\varphi}$ and μ.

5.3 Responsiveness

This section proposes an assortment of properties meant to ensure that changes in the profile produce an intuitive, and expected, change of the outcome, i.e., that the merging operation is responsive to the structure of the profile. Since these properties involve expanding the set $N = \{1, \ldots, n\}$ of agents in the profile we need to make sure that there is a stock of agents on hand if needed to supplement the profile with new elements. We assume, therefore, that the set of agents N who supply formulas to the merging operator is part of some larger subsets, whose elements can be invoked upon request. That being said, we can introduce the following postulates, intended to hold for any profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$, and constraints μ, μ_1, μ_2:

(M$_{\mathsf{MONO}}$) $\Delta_\mu(\vec{\varphi} + \varphi_{n+1}) \wedge \Delta_\mu(\varphi'_{n+1}) \models \Delta_\mu(\vec{\varphi} + \varphi'_{n+1})$.

(M$_{\mathsf{PART}}$) If $\Delta_\mu(\vec{\varphi}) \wedge \varphi_{n+1}$ is consistent, then $\Delta_\mu(\vec{\varphi}) \wedge \varphi_{n+1} \models \Delta_\mu(\vec{\varphi} + \varphi_{n+1})$.

(M$_{\mathsf{RSYM}}$) If $\Delta_\mu(\vec{\varphi})$ is complete and μ has more than one model, then $\Delta_\mu(\varphi_1, \ldots, \varphi_n) \not\models \Delta_\mu(\neg\varphi_1, \ldots, \neg\varphi_n)$.

(M$_{\mathsf{RSVB}}$) If $\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$ is consistent, there is φ_{n+1} such that $\Delta_{\mu_1}(\vec{\varphi} + \varphi_{n+1}) \equiv \Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$.

Postulate M$_{\mathsf{MONO}}$, where 'MONO' stands for *monotonicity*, says that if φ_{n+1} agrees with the profile $\vec{\varphi} + \varphi_{n+1}$ to a certain extent when the constraint is μ, then this agreement is carried over when merging the formulas in the profile $\vec{\varphi} + \varphi'_{n+1}$. Intuitively, the profile $\vec{\varphi} + \varphi'_{n+1}$ can be thought of as being obtained from the profile $\vec{\varphi} + \varphi_{n+1}$ by replacing φ_{n+1} with φ'_{n+1}: it is as if agent $n+1$ considers its options, changes its mind and submits φ'_{n+1} instead of φ_{n+1}. Postulate M$_{\mathsf{MONO}}$ then says that if this change of mind (i.e., after submitting φ'_{n+1}) is in line with the result obtained previously (i.e., when submitting

φ_{n+1}), then the originally agreed upon result should not be changed. In other words, if the agent maintains its support for a raft of issues that were already included in the final result, then these issues are still endorsed by the merging process when the agent submits a formula that expresses as much, if not more, support for these issues. In this, postulate $\mathrm{M}_{\mathsf{MONO}}$ attempts to recreate the monotonicity property found in voting theory: a voting system is monotone if the winning alternative in an election cannot be turned into a non-winner by one voter moving this alternative up in its ranking, while keeping the rest of the ranking fixed. The intuition behind our formalization stems from seeing the models of $\Delta_\mu(\vec{\varphi})$ as the winners in the election where the models of μ are candidates and the formulas in $\vec{\varphi}$ are the voters, and will come out more clearly when modeled as a property for assignments on interpretations, to come shortly. Note that postulate $\mathrm{M}_{\mathsf{MONO}}$ as put forward here is slightly different from the way it was originally presented [Haret et al., 2016b]. The change is made in order to bring the postulate closer to the monotonicity property as featured in social choice. Though arguable whether the present formulation achieves this completely, it is certainly an interesting property to consider.

Postulate $\mathrm{M}_{\mathsf{PART}}$, where '$\mathsf{PART}$' stands for *participation*, refers to a phenomenon that in voting is linked to the *no-show paradox*. A voting rule is vulnerable to this type of paradox if it is possible to change the winner from candidate c_i to candidate c_j by adding a vote in which candidate c_i is strictly preferred to candidate c_j. In a merging scenario, we prevent this by adding a formula φ to a given profile $\vec{\varphi}$ and requiring that $\Delta_\mu(\vec{\varphi}+\varphi)$ should not be 'worse' than $\Delta_\mu(\vec{\varphi})$ with respect to φ.

Postulate $\mathrm{M}_{\mathsf{RSYM}}$, where '$\mathsf{RSYM}$' stands for *reversal symmetry*, harkens back to an eponymous property in voting. A voting rule satisfies reversal symmetry if the winner (assumed to be unique) of an election does not stay a winner if all votes are reversed. In a merging scenario, we interpret the condition of having a unique winner as the outcome of merging being a complete formula (i.e., a formula with exactly one model), and we take reversing the vote to mean that every formula is replaced with its negation.

Postulate $\mathrm{M}_{\mathsf{RSVB}}$, where '$\mathsf{RSVB}$' stands for *resolvability*, says that the output of merging can be refined up to an arbitrary degree by adding just one formula to $\vec{\varphi}$. In a voting scenario resolvability requires that any winner can be made the unique winner by adding a single vote [Tideman, 2006], and postulate $\mathrm{M}_{\mathsf{RSVB}}$ models this intuition.

With the postulates in place, we want to switch now to the semantic view, and see how the postulates are represented at the level off an $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations. Thus, given such an assignment, consider the following properties, expected to hold for any $\mathcal{L}$-profile $\vec{\varphi} = (\varphi_i)_{1\leq i\leq n}$, propositional formulas φ_{n+1}, φ'_{n+1}, sets $\mathcal{W}_1$ and $\mathcal{W}_2$ of interpretations and interpretations w_1 and w_2:

($\mathrm{m}_{\mathsf{MONO}}$) If $w_1 \leq_{\vec{\varphi}_1+\varphi_{n+1}} w_2$ and $w_1 \leq_{\varphi'_{n+1}} w_2$, then $w_1 \leq_{\vec{\varphi}+\varphi'_{n+1}} w_2$.

($\mathrm{m}_{\mathsf{PART}}$) If $w_1 \leq_{\vec{\varphi}} w_2$ and $w_1 \in [\varphi_{n+1}]$, then $w_1 \leq_{\vec{\varphi}+\varphi_{n+1}} w_2$.

($\mathrm{m}_{\mathsf{RSYM}}$) If $w_1 <_{(\varphi_i)_{1\leq i\leq n}} w_2$, then $w_2 <_{(\neg\varphi_i)_{1\leq i\leq n}} w_1$.

($\mathrm{m}_{\mathsf{RSVB}}$) If $w_1 \leq_{\vec{\varphi}} w_2$, for every interpretation $w_1 \in \mathcal{W}_1$ and $w_2 \in \mathcal{W}_2$, then there exists a formula φ_{n+1} such that $w_1 <_{\vec{\varphi}+\varphi_{n+1}} w_2$, for every interpretation $w_1 \in \mathcal{W}_1$ and $w_2 \in \mathcal{W}_2$.

Theorem 5.3

If Δ is an $\mathcal{L}^n$-merging operator satisfying postulates M_{0-1} and M_3, and $\preccurlyeq$ is a total $\mathcal{L}^n$-assignment on interpretations that represents it, then the following equivalences hold:

(1) Δ satisfies postulate $\mathrm{M}_{\mathsf{MONO}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{MONO}}$.

(2) Δ satisfies postulate $\mathrm{M}_{\mathsf{PART}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{PART}}$.

(3) Δ satisfies postulate $\mathrm{M}_{\mathsf{RSYM}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{RSYM}}$.

(4) Δ satisfies postulate $\mathrm{M}_{\mathsf{RSVB}}$ if and only if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{RSVB}}$.

Proof

For Equivalence (1), suppose first that Δ satisfies postulate $\mathrm{M}_{\mathsf{MONO}}$ and take interpretations w_1 and w_2 such that $w_1 \leq_{\vec{\varphi}+\varphi_{n+1}} w_2$ and $w_1 \leq_{\varphi'_{n+1}} w_2$. This implies that $w_1 \in \min_{\leq_{\vec{\varphi}+\varphi_{n+1}}}[\varepsilon_{1,2}]$ and $w_1 \in \min_{\leq_{\varphi'_{n+1}}}[\varepsilon_{1,2}]$, i.e., that $w_1 \in [\Delta_{\varepsilon_{1,2}}(\vec{\varphi}+\varphi_{n+1}) \wedge \Delta_{\varepsilon_{1,2}}(\varphi'_{n+1})]$. Using postulate $\mathrm{M}_{\mathsf{MONO}}$, we conclude that $w_1 \in [\Delta_{\varepsilon_{1,2}}(\vec{\varphi}+\varphi'_{n+1})]$. From this it follows that $w_1 \leq_{\vec{\varphi}+\varphi'_{n+1}} w_2$. Conversely, suppose $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{MONO}}$ and $\Delta_\mu(\vec{\varphi}+\varphi_{n+1}) \wedge \Delta_\mu(\varphi'_{n+1})$ is consistent. The latter fact implies that there exists an interpretation $w_1 \in [\Delta_\mu(\vec{\varphi}+\varphi_{n+1}) \wedge \Delta_\mu(\varphi'_{n+1})]$. Taking an arbitrary interpretation $w_2 \in [\mu]$ and applying property $\mathrm{m}_{\mathsf{MONO}}$, we conclude that $w_1 \leq_{\vec{\varphi}+\varphi'_{n+1}} w_2$, which implies that $w_1 \in [\Delta_\mu(\vec{\varphi}+\varphi'_{n+1})]$.

For Equivalence (2), suppose Δ satisfies postulate $\mathrm{M}_{\mathsf{PART}}$ and take two interpretations w_1 and w_2 such that $w_1 \leq_{\vec{\varphi}} w_2$. We then obtain that $w_1 \in [\Delta_{\varepsilon_{1,2}}(\vec{\varphi}) \wedge \varphi_{n+1}]$, which, by postulate $\mathrm{M}_{\mathsf{PART}}$, implies that $w_1 \in [\Delta_{\varepsilon_{1,2}}(\vec{\varphi}+\varphi_{n+1})]$ and hence $w_1 \leq_{\vec{\varphi}+\varphi_{n+1}} w_2$. Conversely, if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{PART}}$, then for any $w_1 \in [\Delta_\mu(\vec{\varphi}) \wedge \varphi_{n+1}]$ and interpretation $w_2 \in [\mu]$, it follows that $w_1 \leq_{\vec{\varphi}+\varphi_{n+1}} w_2$, which implies the conclusion.

For Equivalence (3), suppose Δ satisfies postulate $\mathrm{M}_{\mathsf{RSYM}}$ and take two distinct interpretations w_1 and w_2 such that $w_1 <_{(\varphi_i)_{1\leq i\leq n}} w_2$. This implies that $[\Delta_{\varepsilon_{1,2}}(\varphi_1,\ldots,\varphi_n)] = \{w_1\}$, i.e., that $\Delta_{\varepsilon_{1,2}}(\varphi_1,\ldots,\varphi_n)$ is complete. Applying postulate $\mathrm{M}_{\mathsf{RSYM}}$, it follows that $[\Delta_{\varepsilon_{1,2}}(\neg\varphi_1,\ldots,\neg\varphi_n)] = \{w_2\}$, showing that property $\mathrm{m}_{\mathsf{RSYM}}$ is satisfied. Conversely, if $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{RSYM}}$ and $[\Delta_\mu(\varphi_1,\ldots,\varphi_n)] = \{w_1\}$, then $w_1 <_{(\varphi_i)_{1\leq i\leq n}} w_2$, for any other interpretation $w_2 \in [\mu]$, which must exist as per the assumption of postulate $\mathrm{M}_{\mathsf{RSYM}}$. Applying property $\mathrm{m}_{\mathsf{RSYM}}$ results in $w_2 <_{(\neg\varphi_i)_{1\leq i\leq n}} w_1$, which delivers the conclusion.

For Equivalence (4), suppose Δ satisfies postulate $\mathrm{M}_{\mathsf{RSVB}}$ and take sets of interpretations $\mathcal{W}_1$ and $\mathcal{W}_2$ such that $w_1 \leq_{\vec{\varphi}} w_2$, for any $w_1 \in \mathcal{W}_1$ and $w_2 \in \mathcal{W}_2$. It follows that $\mathcal{W}_1 \subseteq \min_{\leq_{\vec{\varphi}}}(\mathcal{W}_1 \cup \mathcal{W}_2)$ and hence that $[\Delta_{\varepsilon_{\mathcal{W}_1 \cup \mathcal{W}_2}}(\vec{\varphi}) \wedge \varepsilon_{\mathcal{W}_1}] = \mathcal{W}_1$. Postulate $\mathrm{M}_{\mathsf{RSVB}}$ implies that there exists a formula φ_{n+1} such that $\Delta_{\varepsilon_{\mathcal{W}_1 \cup \mathcal{W}_2}}(\vec{\varphi} + \varphi_{n+1}) \equiv \Delta_{\varepsilon_{\mathcal{W}_1, \mathcal{W}_2}}(\vec{\varphi}) \wedge \varepsilon_{\mathcal{W}_1}$, from which it follows that $[\Delta_{\varepsilon_{\mathcal{W}_1 \cup \mathcal{W}_2}}(\vec{\varphi} + \varphi_{n+1})] = \mathcal{W}_1$, and hence $w_1 <_{\vec{\varphi}+\varphi_{n+1}} w_2$, for any $w_1 \in \mathcal{W}_1$ and $w_2 \in \mathcal{W}_2$. Conversely, suppose $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{RSYM}}$ and take formulas μ_1 and μ_2 such that $\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$ is consistent. This implies that $w_1 \leq_{\vec{\varphi}} w_2$, for every $w_1 \in \min_{\leq_{\vec{\varphi}}}[\mu_1] \cap [\mu_2]$ and $w_2 \in [\mu_1]$ and hence, by property $\mathrm{m}_{\mathsf{RSVB}}$, that there exists a formula φ_{n+1} such that $w_1 <_{\vec{\varphi}+\varphi_{n+1}} w_2$, for every $w_1 \in \min_{\leq_{\vec{\varphi}}}[\mu_1] \cap [\mu_2]$ and $w_2 \in [\mu_1]$. From this it follows that $\Delta_{\mu_1}(\vec{\varphi} + \varphi_{n+1}) \equiv \Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$

Proposition 5.8

If Δ is an $\mathcal{L}^n$-merging operator that satisfies postulates M_0 and M_1, then the following statements hold:

(1) if Δ satisfies postulates M_2 and M_5, then Δ satisfies postulate $\mathrm{M}_{\mathsf{PART}}$.

(2) if Δ satisfies postulates M_2, M_5 and M_6 then Δ satisfies postulate $\mathrm{M}_{\mathsf{RSVB}}$.

Proof

For Statement (1), take $w \in [\Delta_\mu(\vec{\varphi}) \wedge \varphi_{n+1}]$. By postulate M_0 it follows that $w \in [\mu]$. Since $w \in [\varphi_{n+1} \wedge \mu]$, then by postulate M_2 we can conclude that $w \in [\Delta_\mu(\varphi_{n+1})]$ and thus that $w \in [\Delta_\mu(\vec{\varphi}) \wedge \Delta_\mu(\varphi_{n+1})]$. Using postulate M_5 it follows that $w \in [\Delta_\mu(\vec{\varphi} + \varphi_{n+1})]$.

For Statement (2), take $\varphi_{n+1} \equiv \Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$. Both postulates M_0 and M_1 we have that $(\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2) \wedge \mu_1 \equiv \Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$, Thus, using postulate M_2, we have that $\Delta_{\mu_1}(\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2) \equiv \Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$. This shows, among other things, that $\Delta_{\mu_1}(\vec{\varphi}) \wedge \Delta_{\mu_1}(\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2)$ is consistent, which, by postulates M_5 and M_6 implies that $\Delta_{\mu_1}(\vec{\varphi}) \wedge \Delta_{\mu_1}(\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2) \equiv \Delta_{\mu_1}(\vec{\varphi} + (\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2))$. Using the previously derived equivalences, we can conclude that $\Delta_{\mu_1}(\vec{\varphi}) \wedge \Delta_{\mu_1}(\Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2) \equiv \Delta_{\mu_1}(\vec{\varphi}) \wedge \mu_2$.

Before laying down the full picture of how existing merging operators fare with respect to the postulates in this section, a quick observation on the reversal symmetry postulate $\mathrm{M}_{\mathsf{RSYM}}$ will help make things clearer. Reflection on postulate $\mathrm{M}_{\mathsf{RSYM}}$, and even more so on its semantic counterpart, property $\mathrm{m}_{\mathsf{RSYM}}$, shows that the demands it places on an assignment are considerable: in particular, property $\mathrm{m}_{\mathsf{RSYM}}$ requires that replacing all formulas in an $\mathcal{L}$-profile $\vec{\varphi}$ with their negation should reverse all strict comparisons in the preorder corresponding to the negated profile. When coupled with the observation that negating the formulas in an $\mathcal{L}$-profile may create a profile equivalent to the original

	$\vec{\varphi} = (\varphi_1)$ $\{\emptyset, a\}$	(φ_2) $\{abc\}$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi} + \varphi_2, \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximax}}(\vec{\varphi} + \varphi_2, \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximin}}(\vec{\varphi} + \varphi_2, \bullet)$
$\emptyset$	0	3	3	$(3, 0)$	$(0, 3)$
abc	2	0	**2**	$(\mathbf{2}, \mathbf{0})$	$(\mathbf{0}, \mathbf{2})$

Table 5.2: Hamming distances from $\vec{\varphi} = (\varphi_1)$ and φ_2 to each model of μ, with $[\varphi_1] = \{\emptyset, a\}$, $[\varphi_2] = \{abc\}$ and $[\mu] = \{\emptyset, abc\}$, together with the aggregated distances using the sum, leximax and leximin aggregation functions.

vspace1.5em

	$\vec{\varphi} = (\varphi_1)$ $\{\emptyset, a\}$	(φ_2') $\{ab\}$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi} + \varphi_2', \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximax}}(\vec{\varphi} + \varphi_2', \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximin}}(\vec{\varphi} + \varphi_2', \bullet)$
$\emptyset$	0	2	**2**	$(\mathbf{2}, \mathbf{0})$	$(\mathbf{0}, \mathbf{2})$
abc	2	1	3	$(2, 1)$	$(1, 2)$

Table 5.3: Hamming distances from $\vec{\varphi} = (\varphi_1)$ and φ_2 to each model of μ, with $[\varphi_1] = \{\emptyset, a\}$ as above, $[\varphi_2] = \{abc\}$ and $[\mu] = \{\emptyset, abc\}$, together with the aggregated distances using the sum, leximax and leximin aggregation functions.

one, this leads to the conclusion that the only feasible preorder that can represent such a situation is one in which all interpretations are on the same level.

Lemma 5.1

If Δ is an $\mathcal{L}^n$-merging operator that satisfies postulates M_{0-1}, M_3 and $\mathrm{M}_{\mathsf{RSYM}}$, then $\Delta_\top(\varphi, \neg\varphi) \equiv \top$, for any propositional formula φ.

Proof

We know, by Theorem 3.10, that Δ is represented by a total, syntax insensitive and m-faithful $\mathcal{L}^n$-assignment $\preccurlyeq$ on interpretations. By Theorem 5.3, we can also conclude that $\preccurlyeq$ satisfies property $\mathrm{m}_{\mathsf{RSYM}}$. Suppose, now, that there exists a propositional formula φ such that $\Delta_\top(\vec{\varphi}) \not\equiv \top$, where $\vec{\varphi}$ is the $\mathcal{L}$-profile $\vec{\varphi} = (\varphi, \neg\varphi)$. This implies that there exist interpretations w_1 and w_2 such that $w_1 <_{\vec{\varphi}} w_2$. By property $\mathrm{m}_{\mathsf{RSYM}}$, we conclude that $w_2 <_{\vec{\varphi}'} w_1$, where $\vec{\varphi}'$ is the $\mathcal{L}$-profile $\vec{\varphi}' = (\neg\varphi, \neg(\neg\varphi))$. It is easy to see, however, that $\vec{\varphi}$ and $\vec{\varphi}'$ are equivalent profiles and hence, by postulate M_3, that $\Delta_\mu(\vec{\varphi}) \equiv \Delta_\mu(\vec{\varphi}')$, for any propositional formula μ. But this implies that $\leq_{\vec{\varphi}} = \leq_{\vec{\varphi}'}$, which contradicts the conclusions derived previously.

With all these results in hand, we can now have a full picture of how the main merging operators stand up against the responsiveness properties put forward in this section.

	$\vec{\varphi}$	φ_{n+1}	φ'_{n+1}	$d_{\mathrm{D}}^{\mathrm{sum}}(\vec{\varphi}+\varphi_{n+1},\bullet)$	$d_{\mathrm{D}}^{\mathrm{sum}}(\vec{\varphi}+\varphi'_{n+1},\bullet)$
w_1	$k+1$	0	0	$k+1$	$k+1$
w_2	k	1	0	$k+1$	k

Table 5.4: Drastic distances from $\vec{\varphi}$, φ_{n+1} and φ'_{n+1} to w_1 and w_2, together with the aggregated distances using the sum aggregation functions, for a stereotypical case that does not satisfy property $\mathrm{m}_{\mathsf{MONO}}$: outcome w_1 is winning after adding φ_{n+1} to $\vec{\varphi}$, but it loses out to w_2 when agent $n+1$ submits a formula that weakens the support for w_1.

Proposition 5.9

If $\oplus$ is either the sum, leximax or leximin aggregation function, then the following statements hold:

(1) merging operators $\Delta^{\mathrm{H},\oplus}$ and $\Delta^{\mathrm{D},\oplus}$ all satisfy postulates $\mathrm{M}_{\mathsf{PART}}$ and $\mathrm{M}_{\mathsf{RSVB}}$;

(2) merging operators $\Delta^{\mathrm{H},\oplus}$ do not satisfy postulate $\mathrm{M}_{\mathsf{RSYM}}$, but operators $\Delta^{\mathrm{D},\oplus}$ satisfy it;

(3) neither of the merging operators $\Delta^{\mathrm{H},\oplus}$ and $\Delta^{\mathrm{D},\oplus}$ satisfies postulate $\mathrm{M}_{\mathsf{MONO}}$;

Proof

Statement (1) follows from Corollary 3.6, showing that all the operators considered here satisfy postulate M_{0-8} and Proposition 5.8, showing that these postulates guarantee satisfaction of postulates $\mathrm{M}_{\mathsf{PART}}$ and $\mathrm{M}_{\mathsf{RSVB}}$.

For Statement (2) and operators $\Delta^{\mathrm{H},\oplus}$, take the set of atoms $A=\{a,b\}$ and a profile $\vec{\varphi}=(\varphi_1,\varphi_2)$, with $[\varphi_1]=\{\emptyset,b,ab\}$ and $[\varphi_2]=\{a\}$. Notice that $\Delta_{\top}^{\mathrm{H},\oplus}(\vec{\varphi})\not\equiv\top$, for all of the aggregation functions considered and thus, by Lemma 5.1, the merging operators do not satisfy postulate $\mathrm{M}_{\mathsf{RSYM}}$.

For Statement (2) and operators $\Delta^{\mathrm{D},\oplus}$, recall that operators $\Delta^{\mathrm{D},\oplus}$ are equivalent, for all aggregation functions considered here, and that the aggregated distance $d^{\oplus}(\vec{\varphi},w)$ from an $\mathcal{L}$-profile $\vec{\varphi}$ to an interpretation w essentially keeps track of the number of formulas in $\vec{\varphi}$ that have w as their model; obviously, if we replace the formulas in $\vec{\varphi}$ with their negation, then this number is reversed. More precisely, if $\vec{\varphi}'$ is the profile obtained by replacing every formula in $\vec{\varphi}$ with its negation, then $d_{\mathrm{D}}^{\oplus}(\vec{\varphi}',w)=n-d_{\mathrm{D}}^{\oplus}(\vec{\varphi},w)$, where n is the number of agents in the profile. This implies that if $w_1 <_{\vec{\varphi}}^{\mathrm{D},\oplus} w_2$, then $w_2 <_{\vec{\varphi}'}^{\mathrm{D},\oplus} w_1$, for any interpretations w_1 and w_2, which shows that $\preccurlyeq^{\mathrm{D},\oplus}$ satisfies property $\mathrm{m}_{\mathsf{RSYM}}$ and, by Theorem 5.3, postulate $\mathrm{M}_{\mathsf{RSYM}}$ as well.

For Statement (3) and the operators $\Delta^{\mathrm{H},\oplus}$, take the alphabet $A=\{a,b,c\}$, the

d_{H}	$[\varphi_1]$ $\{ab, abc\}$	$[\varphi_2^{\mathsf{t}}]$ $\{ab, ac, abc\}$	$[\varphi_2^{\mathsf{f}}]$ $\{a\}$	$[\varphi_3]$ $\{b\}$	$[\varphi_4]$ $\{c\}$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}^{\mathsf{t}}, \bullet)$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}^{\mathsf{f}}, \bullet)$
ab	0	0	1	1	3	4	5
ac	1	0	1	3	1	5	6
bc	1	1	3	1	1	4	6

Table 5.5: Academy member 2, whose truthful position is expressed by φ_2^{t}, can obtain a better result by submitting φ_2^{f}.

$\mathcal{L}$-profile $\vec{\varphi} = (\varphi_1)$, with $[\varphi_1] = \{\emptyset, a\}$, the propositional formulas φ_2 and φ_2' with $[\varphi_2] = \{abc\}$ and $[\varphi_2'] = \{ab\}$, and a constraint μ with $[\mu] = \{\emptyset, abc\}$. The Hamming distances from $\vec{\varphi} + \varphi_2$ and $\vec{\varphi} + \varphi_2'$ to each model of μ, together with the aggregated distances according to the sum, leximax and leximin aggregation functions are shown in Tables 5.2 and 5.3, respectively. We have that $[\Delta_{\mu}^{\mathrm{H},\oplus}(\vec{\varphi} + \varphi_2)] = \{abc\}$, for all of the aggregation functions considered here, and thus $abc <_{P+\varphi_2}^{\mathrm{H},\oplus} \emptyset$. At the same time, we also have that $abc <_{\varphi_2'}^{\mathrm{H},\oplus} \emptyset$, but $\emptyset <_{\vec{\varphi}+\varphi_2'}^{\mathrm{H},\oplus} abc$.

For Statement (3) and the operators $\Delta^{\mathrm{D},\oplus}$, recall first that operators defined using the aggregation functions considered here are all equivalent, so we make the argument only for $\Delta^{\mathrm{D},\mathrm{sum}}$. Take, now, the alphabet $A = \{a, b\}$, the $\mathcal{L}$-profile $\vec{\varphi} = (\varphi_1)$, with $[\varphi_1] = \{a\}$, the propositional formulas φ_2 and φ_2' with $[\varphi_2] = \{\emptyset\}$ and $[\varphi_2'] = \{\emptyset, a\}$, and a constraint μ with $[\mu] = \{\emptyset, a\}$. We obtain that $\emptyset \approx_{\vec{\varphi}+\varphi_2}^{\mathrm{D},\mathrm{sum}} a$, $\emptyset \approx_{\varphi_2'}^{\mathrm{D},\mathrm{sum}} a$ but $a <_{\vec{\varphi}+\varphi_2'}^{\mathrm{D},\mathrm{sum}} \emptyset$, which constitutes a counterexample to property $\mathrm{m}_{\mathsf{MONO}}$: an edge case, to be sure, but a counterexample nonetheless, the general form of which is depicted in Table 5.4.

5.4 Strategyproofness

In this section we look at issues related to the manipulability and strategyproofness of merging procedures. Issues of strategic reasoning cannot be avoided if, as we have argued, merging is to be used as a framework for collective decision making. A significant concern in any deliberation scenario is that the agents involved may have an incentive to misrepresent their positions, and thus manipulate the aggregation result, if doing so can bring them an advantage. Hence, an understanding of the potential for manipulation of any aggregation procedure is a prerequisite to its successful deployment in any real world context. That merging operators are apt to be manipulated is illustrated by a quick example.

Example 5.6

Recall the example of the four Academy members who have to agree on two nominees for the *Best Director* category, out of three possible directors: Alma Har'el (a), Bong Joon Ho (b) and Céline Sciamma (c). The opinions of the Academy members are $\varphi_1 = a \wedge b$, $\varphi_2 = a \wedge (b \vee c)$, $\varphi_3 = \neg a \wedge b \wedge \neg c$. and $\varphi_4 = \neg a \wedge \neg b \wedge c$, and the constraint is $\mu = (a \wedge b \wedge \neg c) \vee (a \wedge \neg b \wedge c) \vee (\neg a \wedge b \wedge c)$. Suppose, now, that merging is done with the operator $\Delta^{\mathrm{H,\,sum}}$. We saw in Example 3.18 that $[\Delta^{\mathrm{H,\,sum}}_{\mu}(\vec{\varphi})] = \{ab, bc\}$, for $\vec{\varphi} = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$, i.e., the result is to nominate either Alma Har'el and Bong Joon Ho, or Bong Joon Ho and Céline Sciamma. The existence of two possible lineups indicates that the result suggested by the operator $\Delta^{\mathrm{H,\,sum}}$ is not decisive, but is something like a tie between two equally acceptable outcomes. Note, however, that both outcomes agree on b, such that b seems like a safe bet for whatever the final result turns out to be.

Switching our focus to Academy member 2, whose preferences are given by $\varphi_2 = a \wedge (b \vee c)$, we see that they also vacillate between a few options, i.e., ab, ac, abc, but throughout all of them a occurs consistently. We may assume, therefore, that Academy member 2 would prefer an outcome that guarantees that a will be part of it to an outcome that does not.

Suppose, now, that Academy member 2 decides to act strategically and, instead of submitting their true position, which we will henceforth denote by $\varphi_2^{\mathsf{t}} = \varphi_2$, submits the formula $\varphi_2^{\mathsf{f}} = a \wedge \neg b \wedge \neg c$. If we write $\vec{\varphi}^{\mathsf{f}}$ for the profile $\vec{\varphi}^{\mathsf{f}} = (\varphi_1, \varphi_2^{\mathsf{f}}, \varphi_3, \varphi_4)$, then we obtain that $[\Delta^{\mathrm{H,\,sum}}_{\mu}(\vec{\varphi}^{\mathsf{f}})] = \{ab\}$, with the details of this computation spelled out in Table 5.5. This is an outcome that is certainly more appealing to Academy member 2, since it contains the atom a, which figures among all of Academy member 2's most preferred outcomes.

In Example 5.6 we see that one of the agents in the profile has an incentive to misrepresent its true position, since by doing so it can pull the merging result closer to its true opinion. Our purpose in this section will be to formalize the reasoning involved in this type of strategic thinking: we will need a way to quantify what it means for a given result to count as better for an agent than a different result, and analyze the extent to which the primary merging operators are vulnerable to manipulation.

Acceptance notions

As we see in Example 5.6, merging operators may output multiple interpretations, all of which can be seen as winning outcomes. In decision terms, this translates as inconclusiveness with respect to the final verdict. Thus, the set of winning outcomes produced by a merging operator is not always expected to be the final step in a reasoning process: without further means, such a set of interpretations does not give a direct answer to which atoms, i.e., issues, are to be ultimately accepted. One can view the winning set

as a "tie" between all the interpretations in the set. If the decision procedure needs to be explicit about every issue under consideration, then a further reasoning mechanism is required, amounting to a method of breaking ties. To this end, we employ well established acceptance notions from the field of knowledge representation and reasoning: skeptical and credulous consequences [Strasser and Antonelli, 2019].

An *acceptance function* acc is a function $\mathrm{acc}\colon \mathcal{L} \to 2^A$ that maps propositional formulas to sets of atoms in $\vec{\varphi}$. We say that $\mathrm{acc}(\varphi)$ *are the accepted atoms of* φ. For a formula φ, we define the following acceptance notions:

$$\mathrm{skept}(\varphi) = \bigcap_{w \in [\varphi]} w, \qquad \mathrm{cred}(\varphi) = \bigcup_{w \in [\varphi]} w.$$

For a formula φ, an atom is *skeptically accepted* if it is in $\mathrm{skept}(\varphi)$, i.e., if it is true in all models of φ, and *credulously accepted* if it is in $\mathrm{cred}(\varphi)$, i.e., if it is true in at least one model of φ. We will follow established convention in writing the skeptical and credulously accepted atoms as words with the atoms as letters. Skeptical acceptance is equivalent to atom-wise logical entailment, and credulous acceptance indicates support of an atom in at least one model.

Example 5.7: Acceptance notions

In Example 5.6, we obtain that $[\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathfrak{t}})] = \{ab, bc\}$. Thus, it holds that $\mathrm{skept}(\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathfrak{t}})) = b$ and $\mathrm{cred}(\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathfrak{t}})) = abc$.

The acceptance notions introduced here focus on positive literals. Thus, we say that $p \in \mathrm{skept}(\varphi)$ if the atom p is in every model of φ, but we do not treat acceptance of negative literals in a similar fashion, i.e., we are not explicit about atoms that are in none of the models of a formula, and that can thus be thought of as uniformly rejected. This asymmetry is not unusual in a social choice context, where rejection of a candidate is often assimilated to non-acceptance, but would be worth looking at in a more extensive treatment of acceptance notions.

It turns out that there is a duality relation between the indices and aggregation operators defined via skeptical and credulous acceptance that we will want to exploit. Recall that the *dual* $\overline{\varphi}$ *of a formula* φ is obtained by replacing every literal in φ with its negation. If $\vec{\varphi} = (\varphi_i)_{1 \le i \le n}$ is an $\mathcal{L}$-profile, then *the dual* $\overline{\vec{\varphi}}$ *of* $\vec{\varphi}$ is the profile defined as $\overline{\vec{\varphi}} = (\overline{\varphi}_i)_{1 \le i \le n}$. If w is an interpretation, *the dual* $\overline{w}$ *of* w is the complement of w. If $\mathcal{W}$ is a set of interpretations, *the dual* $\overline{\mathcal{W}}$ *of* $\mathcal{W}$ is the set of interpretations defined as $\overline{\mathcal{W}} = \{\overline{w} \mid w \in \mathcal{W}\}$. For a propositional formula φ we have that $\overline{[\varphi]} = [\overline{\varphi}]$.

Proposition 5.10

If $\vec{\varphi}$ is a propositional profile, μ is a constraint, $d \in \{d_{\mathrm{H}}, d_{\mathrm{D}}\}$ is a distance function, and $\oplus \in \{\mathrm{sum}, \mathrm{leximax}, \mathrm{leximin}\}$ is an aggregation function, then it holds that $\overline{\mathrm{skept}(\Delta_{\mu}^{d,\oplus}(\vec{\varphi}))} \equiv \mathrm{cred}(\Delta_{\overline{\mu}}^{d,\oplus}(\overline{\vec{\varphi}}))$.

Proof

It is straightforward to see that $d(w_1, w_2) = d(\overline{w_1}, \overline{w_2})$, for any two interpretations w_1 and w_2 and distance function $d \in \{d_{\mathrm{D}}, d_{\mathrm{H}}\}$. Using this, we can conclude that $\overline{\Delta_{\mu}^{d,\oplus}(\vec{\varphi})} \equiv \Delta_{\overline{\mu}}^{d,\oplus}(\overline{\vec{\varphi}})$. Next, we have that for an atom p, it holds that $p \notin \mathrm{skept}(\Delta_{\mu}^{d,\oplus}(\vec{\varphi}))$ if and only if there exists an interpretation $w \in [\Delta_{\mu}^{d,\oplus}(\vec{\varphi})]$ such that $p \notin w$. Using the previous observation, this is equivalent to $p \in \overline{w}$, for some interpretation $\overline{w} \in [\Delta_{\overline{\mu}}^{d,\oplus}(\overline{\vec{\varphi}})]$, which is in turn equivalent to $p \in \mathrm{cred}(\Delta_{\overline{\mu}}^{d,\oplus}(\overline{\vec{\varphi}}))$.

Proposition 5.10 builds on an interesting symmetry exhibited by the merging operators we work with: the result of merging a profile $\vec{\varphi}$ under a constraint μ and the result of merging $\overline{\vec{\varphi}}$ under constraint $\overline{\mu}$ turn out to be themselves duals of each other. This allows us, once we have found some instance related to the skeptical index, to automatically adapt it to the credulous index.

Example 5.8: Merging and duals

For the set of atoms $A = \{a, b\}$, take a profile $\vec{\varphi} = (\varphi_1, \varphi_2)$, with $\varphi_1 = a \to b$, $\varphi_2 = \neg a$ and $\mu = a$. We obtain $[\Delta_{\mu}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi})] = \{ab\}$, and $\mathrm{skept}(\Delta_{\mu}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi})) = ab$. Taking the duals, we have $\overline{\varphi_1} = \neg a \to \neg b$, $\overline{\varphi_2} = a$ and $\overline{\mu} = \neg a$. Notice that $[\varphi_1] = \{\emptyset, b, ab\}$ and $[\overline{\varphi_1}] = \{ab, a, \emptyset\} = \{\overline{\emptyset}, \overline{b}, \overline{ab}\} = \overline{[\varphi_1]}$, i.e., the models of the dual of φ_1 are the duals of the models of φ_1. We obtain that $[\Delta_{\overline{\mu}}^{\mathrm{H},\,\mathrm{sum}}(\overline{\vec{\varphi}})] = \{\emptyset\}$, which is the same as $\overline{[\Delta_{\mu}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi})]}$ (this equality also holds more generally). Lastly, $\overline{\mathrm{skept}(\Delta_{\mu}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}))} = \mathrm{cred}(\Delta_{\overline{\mu}}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}))$.

Manipulation occurs when an agent, called *the strategic agent*, can influence the merging result in its favor by submitting a formula different from its truthful one. In the following we will typically represent the agent's truthful position by a formula φ^{t}, and the formula with which it manipulates as φ^{f}. We represent the strategic agent's contribution by appending its reported formula to a pre-existing profile $\vec{\varphi}$, with $\vec{\varphi}^{\mathsf{t}} = \vec{\varphi} + \varphi^{\mathsf{t}}$ and $\vec{\varphi}^{\mathsf{f}} = \vec{\varphi} + \varphi^{\mathsf{f}}$ being the truthful and manipulated profiles, respectively. Intuitively, this is as if the strategic agent joins the aggregation process *after* everyone else has submitted their positions. This is merely a notational choice, meant to improve readability, and no generality is lost in this way: since all operators we will look at in this section satisfy the anonymity postulate $\mathrm{M}_{\mathsf{ANON}}$, as presented in Section 5.1, the result never depends on

the order of the formulas in the profile.

Constructive and destructive manipulation with respect to an atom

One of the most basic forms of manipulation is one in which the strategic agent has a specific atom p that it targets for acceptance: the strategic agent may want to see p obtain accepted (or rejected) in the final result. This sets up the stage for the notions we will introduce now and which we call, along the lines of similar concepts from the field of social choice [Conitzer and Walsh, 2016], *constructive* and *destructive* manipulation. A profile $\vec{\varphi}$, constraint μ, distance d, aggregation function $\oplus$ and acceptance notion acc are assumed in most definitions, but, in the interest of concision, are explicitly referred to only under pain of ambiguity. Unless otherwise stated, d ranges over $\{d_{\mathrm{D}}, d_{\mathrm{H}}\}$ and $\oplus$ over $\{\mathrm{sum}, \mathrm{leximax}\}$.

The strategic agent *constructively* acc-*manipulates* $\vec{\varphi}$ *with respect to* p *using* $\varphi^{\mathtt{f}}$ if $p \notin \mathrm{acc}(\Delta_\mu(\vec{\varphi} + \varphi^{\mathtt{t}}))$ and $p \in \mathrm{acc}(\Delta_\mu(\vec{\varphi} + \varphi^{\mathtt{f}}))$, and *destructively* acc-*manipulates* $\vec{\varphi}$ *with respect to* p *using* $\varphi^{\mathtt{f}}$ if $p \in \mathrm{acc}(\Delta_\mu(\vec{\varphi} + \varphi^{\mathtt{t}}))$ and $p \notin \mathrm{acc}(\Delta_\mu(\vec{\varphi} + \varphi^{\mathtt{f}}))$. Intuitively, an agent constructively acc-manipulates with respect to p if it can make p be in the accepted atoms of the aggregation result by submitting $\varphi^{\mathtt{f}}$ instead of $\varphi^{\mathtt{t}}$; similarly, an agent destructively manipulates with respect to p if it can kick p out of the accepted atoms of the result. We say that an operator Δ is acc-*strategyproof* if there is no profile $\vec{\varphi}$, constraint μ, atom p and formulas $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$ s.t. the strategic agent, having $\varphi^{\mathtt{t}}$ as its truthful position, acc-manipulates $\vec{\varphi}$, either constructively or destructively, with respect to p using $\varphi^{\mathtt{f}}$.

We first note that, if $\varphi^{\mathtt{t}}$ is the strategic agent's truthful position, any instance of constructive manipulation with respect to p using $\varphi^{\mathtt{f}}$ is also an instance of destructive manipulation with respect to p, obtained by swapping $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$ as the truthful and manipulating formulas, respectively. Next, our results regarding duality (see Proposition 5.10) imply the following duality for manipulation.

Proposition 5.11

A strategic agent constructively (or destructively) skept-manipulates $\vec{\varphi}$ with respect to p if and only if it destructively (or, respectively, constructively) cred-manipulates $\overline{\vec{\varphi}}$ with respect to p using $\overline{\varphi^{\mathtt{f}}}$, with $\overline{\varphi^{\mathtt{t}}}$ as its truthful position and $\overline{\mu}$ as the constraint.

Proof

Assume an instance of constructive skept-manipulation with respect to p. If $p \notin \mathrm{skept}(\Delta_\mu^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{t}}))$, then $p \in \overline{\mathrm{skept}(\Delta_\mu^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{t}}))}$. Thus, by Proposition 5.10, it holds that $p \in \mathrm{cred}(\Delta_{\overline{\mu}}^{d,\oplus}(\overline{P} + \overline{\varphi^{\mathtt{t}}}))$. Similarly, we get that if $p \in \mathrm{skept}(\Delta_{\overline{\mu}}^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{f}}))$, then $p \notin \mathrm{cred}(\Delta_{\overline{\mu}}^{d,\oplus}(\overline{P} + \overline{\varphi^{\mathtt{f}}}))$. We have obtained, in this way, an instance of destructive cred-manipulation with respect to p.

The proof going from an instance of destructive skept-manipulation to an instance of constructive skept-manipulation with respect to p is entirely analogous.

In other words, an instance of constructive skept-manipulation has a direct counterpart, via the duals, in an instance of destructive cred-manipulation, and likewise for destructive skept-manipulation and constructive cred-manipulation. This simplifies our study as we can focus on only one acceptance notion, with results for the other notion following by Proposition 5.11.

Example 5.9: Constructive skept-manipulation to destructive cred-manipulation

In Example 5.6, Academy member 2 constructively skept-manipulates the profile $\vec{\varphi} = (\varphi_1, \varphi_3, \varphi_4)$ with respect to the atom a, relative to the operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$ and constraint $\mu = (a \wedge b \wedge \neg c) \vee (a \wedge \neg b \wedge c) \vee (\neg a \wedge b \wedge c)$, in that $a \notin \mathrm{skept}(\Delta^{\mathrm{H},\,\mathrm{sum}}_{\mu}(\vec{\varphi}^{\mathtt{t}}))$ but $a \in \mathrm{skept}(\Delta^{\mathrm{H},\,\mathrm{sum}}_{\mu}(\vec{\varphi}^{\mathtt{f}}))$. Consider, now, a merging scenario where every formula is replaced by its dual. In this setting, the truthful position of Academy member 2 is $\overline{\varphi_2^{\mathtt{t}}}$: we obtain that $[\overline{\varphi_2^{\mathtt{t}}}] = \{b, c, bc\}$, the constraint is $\overline{\mu}$, with $[\overline{\mu}] = \{a, b, c\}$, and the profile is $\overline{\vec{\varphi}}$. We obtain that $[\Delta^{\mathrm{H},\,\mathrm{sum}}_{\overline{\mu}}(\overline{\vec{\varphi}^{\mathtt{t}}})] = \{a, c\}$, and $a \in \mathrm{cred}(\Delta^{\mathrm{H},\,\mathrm{sum}}_{\overline{\mu}}(\overline{\vec{\varphi}^{\mathtt{t}}}))$. However, if Academy member 2 now submits $\overline{\varphi_2^{\mathtt{f}}}$, we obtain that $[\Delta^{\mathrm{H},\,\mathrm{sum}}_{\overline{\mu}}(\overline{\vec{\varphi}} + \overline{\varphi_2^{\mathtt{f}}})] = \{c\}$, with $a \notin \mathrm{cred}(\Delta^{\mathrm{H},\,\mathrm{sum}}_{\overline{\mu}}(\overline{\vec{\varphi}} + \overline{\varphi_2^{\mathtt{f}}}))$. Hence, if Academy member 2's truthful position is $\overline{\varphi_2^{\mathtt{t}}}$, then it destructively cred-manipulates $\overline{\vec{\varphi}}$ with respect to a using $\overline{\varphi_2^{\mathtt{f}}}$.

Examples 5.6 and 5.9 already show that the merging operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$ is constructively skept-manipulable (and destructively cred-manipulable). Indeed, Theorem 5.4 shows that this extends to all operators introduced so far. Recall that a formula is complete if it has exactly one model.

Theorem 5.4

For any $n \in \mathbb{N}$ and atom $p \in A$, there exists a profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ and formulas $\varphi^{\mathtt{t}}$, $\varphi^{\mathtt{f}}$ such that the strategic agent constructively (and destructively, respectively) acc-manipulates $\vec{\varphi}$ with respect to p using $\varphi^{\mathtt{f}}$, even if $\mu = \top$ and all φ_i, for $i \in \{1, \ldots, n\}$, as well as $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$, are complete. The instances of manipulation occur relative to all operators $\Delta^{d,\,\oplus}$, for $d \in \{d_{\mathrm{D}}, d_{\mathrm{H}}\}$ and $\oplus \in \{\mathrm{sum}, \mathrm{leximax}, \mathrm{leximin}\}$.

Proof

Without loss of generality, we can assume the target atom p is a. We only showcase the constructive skept-manipulation instances, as corresponding cred-manipulation instances can be obtained using Proposition 5.11 and a destructive manipulation instance can be obtained from a constructive manipulation instance by swapping $\varphi^{\mathtt{t}}$

and $\varphi^{\mathtt{f}}$ as the truthful and manipulating base, respectively, of the strategic agent. We assume, throughout, that $\mu = \top$.

The following argument applies to operators $\Delta^{d,\oplus}$, for $d \in \{d_{\mathrm{D}}, d_{\mathrm{H}}\}$ and $\oplus \in \{\mathrm{sum}, \mathrm{leximax}, \mathrm{leximin}\}$. To obtain constructive skept-manipulation, we take $\varphi^{\mathtt{t}} = \bigwedge_{p\in\vec{\varphi}} \neg p$. Thus, $[\varphi^{\mathtt{t}}] = \{\emptyset\}$ and $\mathrm{skept}(\varphi^{\mathtt{t}}) = \emptyset$. We then do a case analysis depending on whether n is odd or even. In both cases, the agent manipulates using $\varphi^{\mathtt{f}} = a \wedge \bigwedge_{p\in\vec{\varphi}, p\neq a} \neg p$, with $[\varphi^{\mathtt{f}}] = \{a\}$. Each operator is analyzed in turn.

Case 1. If n is even, we write $n = 2k$, for $k \in \mathbb{N}$. For the operators $\Delta^{d,\oplus}$, for $d \in \{d_{\mathrm{D}}, d_{\mathrm{H}}\}$ and $\oplus \in \{\mathrm{sum}, \mathrm{leximax}, \mathrm{leximin}\}$ we take the profile $\vec{\varphi} = (\varphi_1, \ldots, \varphi_{2k})$ such that $[\varphi_1] = \cdots = [\varphi_k] = \{\emptyset\}$ and $[\varphi_{k+1}] = \cdots = [\varphi_{2k}] = \{a\}$. Notice that all bases are complete.

For the operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$, note that in the truthful profile $\vec{\varphi}^{\mathtt{t}} = \vec{\varphi} + \varphi^{\mathtt{t}}$ we have $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}}, \emptyset) = k$ and $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}}, a) = k+1$, while for any other interpretation w we get that $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}}, w) = (\sum_{i=1}^{2k} \delta_i) + \delta^{\mathtt{t}}$, where $\delta_i = d_{\mathrm{H}}^{\mathrm{sum}}(\varphi_i, w)$ and $\delta^{\mathtt{t}} = d_{\mathrm{H}}^{\mathrm{sum}}(\varphi^{\mathtt{t}}, w)$. It is straightforward to see that $\delta_i \geq 1$, for any $i \in \{1, \ldots, 2k\}$ and that $\delta^{\mathtt{t}} \geq 1$ as well. Thus, $\emptyset <_{\vec{\varphi}+\varphi^{\mathtt{t}}}^{\mathrm{H},\,\mathrm{sum}} a$ and $\emptyset <_{\vec{\varphi}+\varphi^{\mathtt{t}}}^{\mathrm{H},\,\mathrm{sum}} w$ for any other interpretation w, i.e., $[\Delta_{\top}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}})] = \{\emptyset\}$. In the manipulated profile $\vec{\varphi}^{\mathtt{f}} = \vec{\varphi} + \varphi^{\mathtt{f}}$ we get that $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi} + \varphi^{\mathtt{f}}, \emptyset) = k+1$ and $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi} + \varphi^{\mathtt{f}}, a) = k$, while for any other interpretation w we get that $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}}, w) = (\sum_{i=1}^{2k} \delta_i) + \delta^{\mathtt{f}}$, where $\delta^{\mathtt{f}} = d_{\mathrm{H}}^{\mathrm{sum}}(\varphi^{\mathtt{f}}, w)$. It is straightforward to see that $\delta^{\mathtt{f}} \geq 1$ and thus $a <_{\vec{\varphi}^{\mathtt{f}}}^{\mathrm{H},\,\mathrm{sum}} \emptyset$ and $a <_{\vec{\varphi}^{\mathtt{f}}}^{\mathrm{H},\,\mathrm{sum}} w$ for any other interpretation w, i.e., $[\Delta_{\top}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathtt{f}})] = \{a\}$. Since $a \notin \mathrm{skept}(\Delta_{\top}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}}))$ but $a \in \mathrm{skept}(\Delta_{\top}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathtt{f}}))$, this counts as an instance of constructive manipulation.

For the operator $\Delta^{\mathrm{H},\,\mathrm{leximax}}$ we reason analogously as for $\Delta^{\mathrm{H},\,\mathrm{sum}}$, and using the same profile $\vec{\varphi}$. Notice that the following equality holds:

$$d_{\mathrm{H}}^{\mathrm{leximax}}(\vec{\varphi}^{\mathtt{t}}, \emptyset) = (\underbrace{1, \ldots, 1}_{k \text{ times}}, \underbrace{0, \ldots, 0}_{(k+1) \text{ times}}),$$

and:

$$d_{\mathrm{H}}^{\mathrm{leximax}}(\vec{\varphi}^{\mathtt{t}}, a) = (\underbrace{1, \ldots, 1}_{(k+1) \text{ times}}, \underbrace{0, \ldots, 0}_{k \text{ times}}),$$

while:

$$d_{\mathrm{H}}^{\mathrm{leximax}}(\vec{\varphi}^{\mathtt{t}}, w) = \mathrm{leximax}(\delta_1, \ldots, \delta_{2k}, \delta^{\mathtt{t}}),$$

for any other interpretation w. It follows then that $[\Delta_{\top}^{\mathrm{H},\,\mathrm{leximax}}(\vec{\varphi}^{\mathtt{t}})] = \{\emptyset\}$, and then that $[\Delta_{\top}^{\mathrm{H},\,\mathrm{leximax}}(\vec{\varphi}^{\mathtt{f}})] = \{a\}$. The argument works for the operator $\Delta^{\mathrm{H},\,\mathrm{leximin}}$ as well and is entirely similar.

For the operators $\Delta^{\mathrm{D},\oplus}$ the argument for $\Delta^{\mathrm{H},\,\mathrm{sum}}$ works here unchanged, since the argument does not rely on the fact that any of the numbers in the vector of distances are greater than 1.

	$[\varphi_1]$ $\{\emptyset\}$	$[\varphi_2]$ $\{\emptyset\}$	$[\varphi_3]$ $\{a\}$	$[\varphi_4]$ $\{a\}$	$[\varphi^{\mathtt{t}}]$ $\{\emptyset\}$	$[\varphi^{\mathtt{f}}]$ $\{a\}$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}}, \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximax}}(\vec{\varphi}^{\mathtt{t}}, \bullet)$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}^{\mathtt{f}}, \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximax}}(\vec{\varphi}^{\mathtt{f}}, \bullet)$
$\emptyset$	0	0	1	1	0	1	2	$(1,1,0,0,0)$	3	$(1,1,1,0,0)$
a	1	1	0	0	1	0	3	$(1,1,1,0,0)$	2	$(1,1,0,0,0)$
b	1	1	2	2	1	2	7	$(2,2,1,1,1)$	8	$(2,2,2,1,1)$
...	...	...	...	...	...	...	...	...	...	...

Table 5.6: Constructive skept-manipulation of a profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq 4}$ with respect to the atom a, relative to the operators $\Delta^{\mathrm{H,\,sum}}$ and $\Delta^{\mathrm{H,\,leximax}}$.

Case 2. If n is odd, we write $n = 2k + 1$, for $k \in \mathbb{N}$. For the operators $\Delta^{d,\oplus}$, for $d \in \{d_{\mathrm{D}}, d_{\mathrm{H}}\}$ and $\oplus \in \{\mathrm{sum}, \mathrm{leximax}\}$ we take the profile $\vec{\varphi} = (\varphi_1, \ldots, \varphi_{2k+1})$ such that $[\varphi_1] = \cdots = [\varphi_k] = \{\emptyset\}$ and $[\varphi_{k+1}] = \cdots = [\varphi_{2k+1}] = \{a\}$. Notice that all bases are complete. Calculation of the scores for the interpretations, while not completely analogous to the previous case, is sufficiently similar to yield the conclusion.

Theorem 5.4 suggests that the situation with respect to constructive and destructive manipulation is acute, for two reasons. Firstly, restrictions on the size of the profile or on the specificity of the formulas (e.g., requiring that all formulas are complete), which ensure strategyproofness in other contexts [Everaere et al., 2007], turn out not to have any effect in this case. Second, instances of manipulation exist for *any* size of the profile $\vec{\varphi}$: this is best understood by consulting Example 5.10 below.

Example 5.10: Manipulation with respect to an atom

To constructively skept-manipulate a profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq 4}$ with respect to the atom a, relative to the constraint $\mu = \top$ and $\oplus \in \{\mathrm{sum}, \mathrm{leximax}\}$, take φ_i, for $i \in \{1, 2, 3, 4\}$, $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$ as in Table 5.6. It is straightforward to see that $[\Delta_{\mu}^{d,\oplus}(\vec{\varphi}^{\mathtt{t}})] = \{\emptyset\}$ and $[\Delta_{\mu}^{d,\oplus}(\vec{\varphi}^{\mathtt{f}})] = \{a\}$, for $d \in \{d_{\mathrm{D}}, d_{\mathrm{H}}\}$ and $\oplus \in \{\mathrm{sum}, \mathrm{leximax}\}$. Table 5.6 shows results for d_{H}, but the reasoning for d_{D} is entirely similar. This example easily generalizes to any even n. If n is odd, which we can write as $n = 2p + 1$, for $p \in \mathbb{N}$, we can take $[\varphi_1] = \cdots = [\varphi_p] = \{\emptyset\}$, $[\varphi_{p+1}] = \cdots = [\varphi_n] = \{a\}$, and $\varphi^{\mathtt{t}}$, $\varphi^{\mathtt{f}}$ as above.

If possible for an agent to constructively or destructively manipulate, it is appropriate to ask *how* it can do it: are intricate formulas needed to achieve the goal, or can a 'simple' formula work just as well? In Example 5.10 the strategic agent manipulates using complete formulas, suggesting that the answer lies with the second option. Indeed we can show that, if manipulation is possible at all, then it can be done with a complete formula. Before presenting the main result, however, we introduce some helping lemmas.

The first lemma says that if an interpretation w_1 is considered better than w_2 by a formula φ, then w_1 is considered better than w_2 also by the formula φ_*, where φ_* is a complete formula whose model is the model of φ closest to w_1 among all the models of φ.

	$[\varphi]$				$[\varphi_*]$	
	v_1	$\dots$	v_m	$d(\varphi, \bullet)$	v_*	$d(\varphi_*, \bullet)$
w_1	$\delta_{1,1}$	$\dots$	$\delta_{m,1}$	$\delta_{\min,1}$	$\delta_{\min,1}$	$\delta_{\min,1}$
w_2	$\delta_{1,2}$	$\dots$	$\delta_{m,2}$	$\delta_{\min,2}$	$\delta_{\min,2} + \epsilon$	$\delta_{\min,2} + \epsilon$

Table 5.7: Replacing φ with φ_*, where $[\varphi_*] = \{v_*\}$ and v_* is the model of φ closest to w_1, preserves the order between w_1 and w_2, i.e., if w_1 is as good as w_2 relative to φ, then w_1 is also as good as w_2 relative to φ_*.

Lemma 5.2

If φ is a formula, $d \in \{d_{\mathrm{H}}, d_{\mathrm{D}}\}$, w_1 and w_2 are two interpretations and φ_* is a complete formula whose model v_* is such that $v_* \in [\varphi]$ and $d(v_*, w_1) = \min(d(v, w_1))_{v \in [\varphi]}$, then it holds that:

(*i*) if $w_1 <_{\varphi}^{d} w_2$, then $w_1 <_{\varphi_*}^{d} w_2$;

(*ii*) if $w_1 \approx_{\varphi}^{d} w_2$, then $w_1 \leq_{\varphi_*}^{d} w_2$.

Proof

We write $[\varphi] = \{v_1, \dots, v_m\}$ and $d(v_j, w_k) = \delta_{j,k}$, for $k \in \{1, 2\}$, $j \in \{1, \dots, m\}$. Additionally, we write $\delta_{\min,k} = \min(\delta_{1,k}, \dots, \delta_{m,k})$, for $k \in \{1, 2\}$. Table 5.7 illustrates these notions. By definition, we have that $d(\varphi, w_k) = \delta_{\min,k}$, for $k \in \{1, 2\}$.

We start with Claim (*i*): by assumption, it holds that $\delta_{\min,1} < \delta_{\min,2}$. We take now an interpretation $v_* \in [\varphi]$ that is closest to w_1 among the models of φ, i.e., $d(v_*, w_1) = \min(d(v, w_1))_{v \in [\varphi]}$, and a formula φ_* such that $[\varphi_*] = \{v_*\}$. There might be more than one interpretation that is equidistant to w_1 and fits this description, in which case we pick one at random. Switching our attention to the preorder $\leq_{\varphi_*}^{d}$, we have, by definition, that $d(\varphi_*, w_1) = \min(d(v, w_1))_{v \in [\varphi_*]}$, which implies that $d(\varphi_*, w_1) = \delta_{*,1} = \delta_{\min,1}$. At the same time, it holds that $d(\varphi_*, w_2) = \delta_{*,2} = \delta_{\min,2} + \epsilon$, for some $\epsilon \geq 0$. The latter claim is just a rewriting of the fact that $\delta_{*,2} \geq \delta_{\min,2}$, and it follows from the fact that $\delta_{\min,2} = \min(\delta_{1,2}, \dots, \delta_{*,2}, \dots, \delta_{m,2})$. Since, by assumption, $\delta_{\min,1} < \delta_{\min,2}$, then it also holds that $\delta_{\min,1} < \delta_{\min,2} + \epsilon$, and hence $d(\varphi_*, w_1) < d(\varphi_*, w_2)$. For Claim (*ii*), our assumption is equivalent to the fact $\delta_{\min,1} = \delta_{\min,2}$, from which it follows that $\delta_{\min,1} \leq \delta_{\min,2} + \epsilon$ and hence $d(\varphi_*, w_1) \leq d(\varphi_*, w_2)$.

Note that ϵ, here, denotes a positive real number, and should not be confused with the proxy $\varepsilon_{\mathcal{W}}$ of some set of interpretations $\mathcal{W}$: even though the symbols are similar, they are nonetheless different, and the entities they refer to are definitely different.

For the next step, we want to recreate the conclusion of Lemma 5.2 in the presence of an

	$\vec{\varphi}$	$[\varphi]$ $\{v_1, \ldots, v_m\}$	$[\varphi_*]$ $\{v_*\}$	$d^{\text{sum}}(\vec{\varphi}+\varphi, \bullet)$	$d^{\text{sum}}(\vec{\varphi}+\varphi_*, \bullet)$
w_1	β_1	$\delta_{\text{min},1}$	$\delta_{\text{min},1}$	$\beta_1 + \delta_{\text{min},1}$	$\beta_1 + \delta_{\text{min},1}$
w_2	β_2	$\delta_{\text{min},2}$	$\delta_{\text{min},2} + \epsilon$	$\beta_2 + \delta_{\text{min},2}$	$\beta_2 + \delta_{\text{min},2} + \epsilon$

Table 5.8: Replacing φ with φ_* in the profile $\vec{\varphi} + \varphi$, where $[\varphi_*] = \{v_*\}$ and v_* is the model of φ closest to w_1, preserves the order between w_1 and w_2, i.e., if w_1 is as good as w_2 relative to the profile $\vec{\varphi} + \varphi$, then w_1 is also as good as w_2 relative to the profle $\vec{\varphi} + \varphi_*$.

aggregation function, i.e., we want to show that order between two interpretations w_1 and w_2 relative to a profile $\vec{\varphi} + \varphi$ is preserved when replacing φ with a carefully selected complete formula φ_*: as for Lemma 5.2, the formula φ_* is a formula based on φ that maximizes the support for w_1, i.e., whose model is a model of φ that is closest to w_1 according to the distance function used.

Lemma 5.3

If $\vec{\varphi}$ is an $\mathcal{L}$-profile, φ is a propositional formula, $d \in \{d_{\text{H}}, d_{\text{D}}\}$ is a distance function, $\oplus \in \{\text{sum}, \text{leximax}, \text{leximin}\}$ is an aggregation function, w_1 and w_2 are two interpretations and φ_* is a complete formula whose model v_* is such that $v_* \in [\varphi]$ and $d(v_*, w_1) = \min(d(v, w_1))_{v \in [\varphi]}$, then it holds that:

(*i*) if $w_1 <^{d,\oplus}_{\vec{\varphi}+\varphi} w_2$, then $w_1 <^{d,\oplus}_{\vec{\varphi}+\varphi_*} w_2$;

(*ii*) if $w_1 \approx^{d,\oplus}_{\vec{\varphi}+\varphi} w_2$, then $w_1 \leq^{d,\oplus}_{\vec{\varphi}+\varphi_*} w_2$;

Proof

We first show the claim for the sum aggregation function, as it provides a nice illustration of the main ideas. For this, we write $d^{\text{sum}}(\vec{\varphi}, w_k) = \beta_k$, for $k \in \{1, 2\}$. Assuming that $[\varphi] = \{v_1, \ldots, v_m\}$, we write $\min(d(v, w_k))_{v \in [\varphi]} = \delta_{\text{min},k}$, for $k \in \{1, 2\}$ Table 5.8 provides an illustration of the main notions used here. By definition, $d(\varphi, w_k) = \delta_{\text{min},k}$, for $k \in \{1, 2\}$. We take now an interpretation $v_* \in [\varphi]$ that is closest to w_1 among the models of φ, i.e., $d(v_*, w_1) = \min(d(v, w_1))_{v \in [\varphi]}$, and a base φ_* such that $[\varphi_*] = \{v_*\}$. We now have that $d(\varphi_*, w_1) = \delta_{*,1} = \delta_{\text{min},1}$, while $d(\varphi_*, w_2) = \delta_{*,2} = \delta_{\text{min},2} + \epsilon$, for some $\epsilon \geq 0$ where the quanitites used here are defined as in Lemma 5.2. We obtain that $d^{\text{sum}}(\vec{\varphi} + \varphi, w_1) = \beta_1 + \delta_{\text{min},1}$ and $d^{\text{sum}}(\vec{\varphi}+\varphi, w_2) = \beta_2+\delta_{\text{min},2}$. Additionally, we have that $d^{\text{sum}}(\vec{\varphi}+\varphi_*, w_1) = \beta_1+\delta_{\text{min},1}$ and $d^{\text{sum}}(\vec{\varphi}+\varphi_*, w_2) = \beta_2+\delta_{\text{min},2}+\epsilon$. If $\beta_1+\delta_{\text{min},1} < \beta_2+\delta_{\text{min},2}$, as per the assumption of Claim (*i*), then $\beta_1 + \delta_{\text{min},1} < \beta_2 + \delta_{\text{min},2} + \epsilon$ and hence $d^{\text{sum}}(\vec{\varphi} + \varphi_*, w_1) < d^{\text{sum}}(\vec{\varphi} + \varphi_*, w_2)$. If $\beta_1 + \delta_{\text{min},1} = \beta_2 + \delta_{\text{min},2}$, as per the assumption of Claim (*ii*),

then $\beta_1 + \delta_{\min,1} \leq \beta_2 + \delta_{\min,2} + \epsilon$ and hence $d^{\text{sum}}(\vec{\varphi} + \varphi_*, w_1) \leq d^{\text{sum}}(\vec{\varphi} + \varphi_*, w_2)$.

For the leximax aggregation function, the argument has to be adapted to the output for each aggregation function, but is otherwise entirely similar. The integers β_1 and β_2 (i.e., the distances from $\vec{\varphi}$ to w_1 and w_2) must be replaced with tuples of integers $B_1 = (\beta_{1,1}, \beta_{2,1}, \dots)$ and $B_2 = (\beta_{1,2}, \beta_{2,2}, \dots)$. For Claim (i) we then have, by assumption, that $\text{leximax}(\beta_{1,1}, \beta_{2,1}, \dots, \delta_{\min,1}) <_{\text{lex}} \text{leximax}(\beta_{1,2}, \beta_{2,2}, \dots, \delta_{\min,2})$. Since $\delta_{*,2} \leq \delta_{*,2} + \epsilon$ and leximax satisfies the monotonicity property Ag_3 of aggregation functions, presented in Section 2.3, we obtain that $\text{leximax}(\beta_{1,1}, \beta_{2,1}, \dots, \delta_{*,1}) <_{\text{lex}} \text{leximax}(\beta_{1,2}, \beta_{2,2}, \dots, \delta_{*,2} + \epsilon)$ and thus $d^{\text{leximax}}(\vec{\varphi} + \varphi_*, w_1) \leq d^{\text{leximax}}(\vec{\varphi} + \varphi_*, w_2)$. The argument for Claim (ii) is entirely similar, and the proof for the leximin aggregation function follows analogously.

We can now use Lemma 5.3 to show that if manipulation with respect to an atom is possible, then it is possible to manipulate using a complete base.

Theorem 5.5

If the strategic agent constructively, or destructively, acc-manipulates $\vec{\varphi}$ with respect to p using $\varphi^{\mathtt{f}}$, for $\text{acc} \in \{\text{skept}, \text{cred}\}$, then there exists a complete formula $\varphi^{\mathtt{f}}_*$ such that $\varphi^{\mathtt{f}}_* \models \varphi^{\mathtt{f}}$ and the agent constructively, or destructively, skept-manipulates $\vec{\varphi}$ with respect to p using $\varphi^{\mathtt{f}}_*$.

Proof

We prove the claim for constructive skept-manipulation first. The fact that the strategic agent skept-manipulates $\vec{\varphi}$ using $\varphi^{\mathtt{f}}$ implies that there exist interpretations $w_1, \dots, w_l$ in $[\mu]$ such that $p \in \text{skept}(\{w_1, \dots, w_l\})$, and $[\Delta_\mu^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{f}})] = \{w_1 \dots, w_l\}$. We pick one of the interpretations in $[\Delta_\mu^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{f}})]$, say w_1. Take, now, $v_* \in [\varphi^{\mathtt{f}}]$ such that $d(v_*, w_1) = \min(d(v, w_1))_{v \in [\varphi^{\mathtt{f}}]}$, i.e., a model of $\varphi^{\mathtt{f}}$ that is closest to w_1. The claim now is that we can constructively skept-manipulate $\vec{\varphi}$ with $\varphi^{\mathtt{f}}_*$, where $[\varphi^{\mathtt{f}}_*] = \{v_*\}$. This follows by observing that $w_1 \leq^{d,\oplus}_{\vec{\varphi} + \varphi^{\mathtt{f}}} w_i$, for all $w_i \in [\mu]$ and thus, by Lemma 5.3, it follows that $w_1 \leq^{d,\oplus}_{\vec{\varphi} + \varphi^{\mathtt{f}}_*} w_i$, for all $w_i \in [\mu]$. Thus, w_1 stays part of the aggregation result. Additionally, if $w_1 <^{d,\oplus}_{\vec{\varphi} + \varphi^{\mathtt{f}}} w_i$, for some $w_i \in [\mu]$, then, again by Lemma 5.3, it follows that $w_1 <^{d,\oplus}_{\vec{\varphi} + \varphi^{\mathtt{f}}_*} w_i$. In summary, by replacing $\varphi^{\mathtt{f}}$ with $\varphi^{\mathtt{f}}_*$, w_1 and possibly some other interpretations in $\{w_1, \dots, w_l\}$ remain winning, and no new winning interpretations are added. Another way of putting this is that $[\Delta_\mu^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{f}}_*)] \subseteq [\Delta_\mu^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{f}})]$. Since $p \in \text{skept}(\Delta_\mu^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{f}}))$, we get that $p \in \text{skept}(\Delta_\mu^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{f}}_*))$ as well.

For destructive skept-manipulation, we get that there exists an interpretation $w_1 \in [\Delta_\mu^{d,\oplus}(\vec{\varphi} + \varphi^{\mathtt{f}})]$ such that $p \notin w_1$. We pick, as before, a model v_* of φ that is closest to

w_1 among the models of φ, i.e., $d(v_*, w_1) = \min(d(v, w_1))_{v \in [\varphi^{\mathtt{f}}]}$. Using Lemma 5.3, we again obtain that $w_1 \in [\Delta_\mu^{d, \oplus}(\vec{\varphi} + \varphi_*^{\mathtt{f}})]$. This guarantees that $p \notin \mathrm{skept}(\Delta_\mu^{d, \oplus}(\vec{\varphi} + \varphi_*^{\mathtt{f}}))$.

The case for constructive, or destructive, cred-manipulation follows by applying Proposition 5.11.

The intuition driving the proof for skept-manipulation in Theorem 5.5 is that if manipulation is possible with $\varphi^{\mathtt{f}}$, then we pick a model of $\varphi^{\mathtt{f}}$ that is closest to one of the models of μ crucial for the success of manipulation. In the case of destructive skept-manipulation, this would be an interpretation v_* that ends up being in $[\Delta_\mu^{d, \oplus}(\vec{\varphi} + \varphi^{\mathtt{f}})]$ and is such that $p \notin v_*$: v_* must exist, under the assumption that $\varphi^{\mathtt{f}}$ successfully achieves destructive skept-manipulation. We can then replace $\varphi^{\mathtt{f}}$ with $\varphi_*^{\mathtt{f}}$, where $[\varphi_*^{\mathtt{f}}] = \{v_*\}$ and still achieve destructive skept-manipulation.

There is one thing that mitigates the acuteness of the manipulation results. Note that we have not assumed so far that the strategic agent needs to have p among its accepted atoms, i.e., we do not require the agent to actually *believe* p in order to constructively/destructively manipulate with respect to it. Seeing the merging process as aggregating agents' *reported* beliefs, comes into play, as it allows for agents to participate with formulas that can reflect a richer cognitive structure (e.g., the effects of bribery, or influence, motivating an agent to alter its reported beliefs). Thus, here we operate under the assumption that p (its acceptance, or otherwise) figures for the agent as a goal, regardless of whether it is actually part of its beliefs. Manipulation furthering the truthful beliefs of the strategic agent will be touched on shortly.

Can an agent influence the acceptance of an atom it does not, strictly speaking, believe? The answer, in general, is yes: a strategic agent can constructively skept-manipulate with respect to an atom p even though p is not among the skeptical beliefs of the agent itself. And, in fact, we are able to show that, when $\mu = \top$ and all formulas are complete, skept-manipulation is possible only under this assumption.

Proposition 5.12

If the strategic agent constructively skept-manipulates $\vec{\varphi}$ with respect to an atom p, relative to the constraint $\mu = \top$ and operator $\Delta^{\mathrm{H,\,sum}}$, when all formulas are complete, then $p \notin \mathrm{skept}(\varphi^{\mathtt{t}})$.

Proof

The operator $\Delta^{\mathrm{H,\,sum}}$, for complete bases and $\mu = \top$ acts as a majority operator. In other words: if an atom p is accepted by a majority of the agents, then p is in the result; if p is not accepted by a majority of the agents, then p is not in the result; and if there is equality with respect to acceptance of p, then the result features a model that

contains p and a model that does not. This being said, if an agent can constructively skept-manipulate with respect to atom p, then this means, by definition, that p is not in $\mathrm{skept}(\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\varphi^{\mathtt{t}}))$, but that $p \in \mathrm{skept}(\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\varphi^{\mathtt{f}}))$. This implies that the strategic agent's influence over the result consists in inducing a majority for p: the result (with the base of the strategic agent) goes from being undecided with respect to p (and hence p not being skeptically accepted in the result) when the strategic agent is honest, to being in favor of p when the strategic agent submits a base different from its truthful on. In this, the strategic agent is the decisive agent who tips the balance in favor of p: but this can only happen if p is not in $\mathrm{skept}(\varphi^{\mathtt{t}})$ to begin with.

Proposition 5.12 can be seen as a positive result, one way of reading it being that if the strategic agent already accepts p, i.e., $p \in \mathrm{skept}(\varphi^{\mathtt{t}})$, then if it cannot impose p by submitting $\varphi^{\mathtt{t}}$ itself, for the given parameters, then there is no other way of doing it. As such, this is the closest we can come to a strategyproofness result for constructive and destructive manipulation with respect to an atom.

Manipulation with respect to a dissatisfaction index

Constructive and destructive manipulation deals with the question of whether an agent can affect the acceptance of an atom in the aggregated outcome, regardless of the beliefs of the agent. In this section we look at the case when the agent improves the outcome with respect to its true belief. To make sense of this notion of improvement, we need to be able to measure an agent's satisfaction with respect to the result of a merging operator. To this end we introduce a set of dissatisfaction indices that build on the acceptance notions. A *dissatisfaction index* i is a function $i\colon \mathcal{L} \times \mathcal{L} \to \mathbb{N}^+$ that maps a pair of formulas to a non-negative integer [Everaere et al., 2007]. If φ_1 and φ_2 are two propositional formulas and acc is an acceptance notion, *the dissatisfaction index* i_{acc} is defined as:

$$i_{\mathrm{acc}}(\varphi_1, \varphi_2) = d_{\mathrm{H}}(\mathrm{acc}(\varphi_1), \mathrm{acc}(\varphi_2)),$$

i.e., as the number of atoms on which $\mathrm{acc}(\varphi_1)$ and $\mathrm{acc}(\varphi_2)$ differ. For the two acceptance notions introduced above, this gives us the dissatisfaction indices i_{skept} and i_{cred}.

Example 5.11: Dissatisfaction indices

For Academy member 2 in Example 5.6, we have that their truthful opinion is given by $[\varphi_2^{\mathtt{t}}] = \{a, ab, ac\}$. Hence their skeptically accepted atoms are $\mathrm{skept}(\varphi_2^{\mathtt{t}}) = a$. The result of merging when Academy member 2 submits its truthful opinion is $[\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}})] = \{ab, bc\}$, with the skeptically accepted atoms of this result being $\mathrm{skept}(\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}})) = b$. Thus, according to the skeptical dissatisfaction index, we have that $i_{\mathrm{skept}}(\varphi_2^{\mathtt{t}}, \Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}^{\mathtt{t}})) = d_{\mathrm{H}}(a, b) = 2$, giving Academy member 2's level of satisfaction with the truthful result of merging.

For arbitrary formulas the numeric results given by the indices i_{skept} and i_{cred} are generally not directly correlated, in that each may be higher or lower than the other.

A strategic agent whose truthful beliefs are $\varphi^{\mathtt{t}}$ *manipulates* $\vec{\varphi}$ *with respect to* i_{acc} *using* $\varphi^{\mathtt{f}}$ if it holds that $i_{\text{acc}}(\varphi^{\mathtt{t}}, \Delta_\mu(\vec{\varphi} + \varphi^{\mathtt{f}})) < i_{\text{acc}}(\varphi^{\mathtt{t}}, \Delta_\mu(\vec{\varphi} + \varphi^{\mathtt{t}}))$. In other words, the strategic agent manipulates with respect to i_{acc} if it can improve its dissatisfaction index by submitting $\varphi^{\mathtt{f}}$ instead of $\varphi^{\mathtt{t}}$. We say that an operator Δ is *strategyproof with respect to a dissatisfaction index* i_{acc} if there is no profile $\vec{\varphi}$, constraint μ and formulas $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$ such that the strategic agent, having $\varphi^{\mathtt{t}}$ as its truthful position, manipulates $\vec{\varphi}$ with respect to i_{acc} using $\varphi^{\mathtt{f}}$.

Our definition of manipulability based on satisfaction indices is inspired by previous work on manipulation of propositional merging operators [Everaere et al., 2007] but differs from it in an important respect: we measure the distance between the *accepted* atoms of the manipulating agent and the result, rather than between the sets of models themselves.

Example 5.12: Manipulation with respect to a dissatisfaction index

We have seen, in Example 5.11, that for the setting in Example 5.6, we obtain that $i_{\text{skept}}(\varphi_2^{\mathtt{t}}, \Delta_\mu^{\text{H, sum}}(\vec{\varphi}^{\mathtt{t}})) = d_{\text{H}}(a, b) = 2$. In other words, the dissatisfaction of Academy member 2 and the result of merging when Academy member 2 submits their true opinions, according to the skeptical dissatisfaction index, is 2. We have also seen, in Example 5.6, that by changing its reported opinion to $\varphi_2^{\mathtt{f}}$, Academy member 2 is able to change the result to $[\Delta_\mu^{\text{H, sum}}(\vec{\varphi}^{\mathtt{f}}))] = \{ab\}$. It holds, then, that $\text{skept}(\Delta_\mu^{\text{H, sum}}(\vec{\varphi}^{\mathtt{f}}))) = ab$, and hence $i_{\text{skept}}(\varphi_2^{\mathtt{t}}, \Delta_\mu^{\text{H, sum}}(\vec{\varphi}^{\mathtt{f}})) = d_{\text{H}}(a, ab) = 1$. Thus, by submitting a position different from its truthful one, Academy member 2 is able to bring the (skeptically accepted atoms of) the merging result closer to its own position.

Example 5.12 shows that manipulation is possible in the general case for the merging operator $\Delta^{\text{H, sum}}$ and the skeptical index. What is, now, the full picture with respect to manipulability?

As for constructive and destructive manipulation, we first note that there is a duality between the skeptical dissatisfaction notion and the credulous one, given by the identity $i_{\text{skept}}(\varphi_1, \varphi_2) = i_{\text{cred}}(\overline{\varphi_1}, \overline{\varphi_2})$. Intuitively, the identity holds because an atom p being in the symmetric difference of the skeptical consequences is equivalent to there being a model of one of the formulas not containing p, with the dual having p in at least one model. This identity allows us to turn a manipulation instance with respect to i_{skept} into a manipulation instance with respect to i_{cred} simply by replacing every formula involved with its dual.

For the operators $\Delta^{d, \text{leximax}}$ and $\Delta^{d, \text{leximin}}$ index manipulation turns out to be, like atom manipulation, unavoidable. This stays so even under heavy restrictions (i.e., complete formulas and $\mu = \top$), and for any size $n \geq 2$ of the profile.

Theorem 5.6

For $d \in \{d_{\mathrm{D}}, d_{\mathrm{H}}\}$, $\oplus \in \{\mathrm{leximax}, \mathrm{leximin}\}$ and any $n \geq 2$ there exists a profile $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$ and formulas $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$ such that the strategic agent manipulates $\vec{\varphi}$ with respect to i_{acc}, even if $\mu = \top$ and all formulas φ_i, for $i \in \{1, \ldots, n\}$, as well as $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$, are complete.

Proof

We showcase here instances of manipulation with respect to i_{skept} for a profile $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$ of size $n \geq 2$. Instances of manipulation with respect to i_{cred} are obtained by taking the duals of all formulas involved in the instances of manipulation with respect to i_{skept}. We assume that $\mu = \top$.

For the operator $\Delta^{\mathrm{H},\,\mathrm{leximax}}$, take $[\varphi_i] = \{a\}$, for $i \in \{1, \ldots, n\}$, $[\varphi^{\mathtt{t}}] = \{\emptyset\}$ and $[\varphi^{\mathtt{f}}] = \{b\}$. We obtain that $[\Delta_{\mu}^{\mathrm{H},\,\mathrm{leximax}}(\vec{\varphi} + \varphi^{\mathtt{t}})] = \{a\}$ and $[\Delta_{\mu}^{\mathrm{H},\,\mathrm{leximax}}(\vec{\varphi} + \varphi^{\mathtt{f}})] = \{\emptyset, ab\}$. Hence, $i_{\mathrm{skept}}(\varphi^{\mathtt{t}}, \Delta_{\mu}^{\mathrm{H},\,\mathrm{leximax}}(\vec{\varphi} + \varphi^{\mathtt{t}})) = d_{\mathrm{H}}(\emptyset, a) = 1$, while $i_{\mathrm{skept}}(\varphi^{\mathtt{t}}, \Delta_{\mu}^{\mathrm{H},\,\mathrm{leximax}}(\vec{\varphi} + \varphi^{\mathtt{f}})) = d_{\mathrm{H}}(\emptyset, \emptyset) = 0$.

For the operator $\Delta^{\mathrm{H},\,\mathrm{leximin}}$, take $[\varphi_1] = \{a\}$, $[\varphi_i] = \{b\}$, for $i > 1$, $[\varphi^{\mathtt{t}}] = \{b\}$ and $[\varphi^{\mathtt{f}}] = \{\emptyset, ab\}$. We obtain that $[\Delta_{\mu}^{\mathrm{H},\,\mathrm{leximin}}(\vec{\varphi} + \varphi^{\mathtt{t}})] = \{a\}$ and $[\Delta_{\mu}^{\mathrm{H},\,\mathrm{leximin}}(\vec{\varphi} + \varphi^{\mathtt{f}})] = \{\emptyset, a, ab\}$.

For the operator $\Delta^{\mathrm{D},\,\mathrm{leximax}}$ we make a distinction according to whether n is odd or even. If $n = 2k$, take $[\varphi_1] = \cdots = [\varphi_k] = \{\emptyset\}$, $[\varphi_{k+1}] = \cdots = [\varphi_{2k}] = \{a\}$, $[\varphi^{\mathtt{t}}] = \{ab\}$ and $[\varphi^{\mathtt{f}}] = \{a\}$. We get that $[\Delta_{\mu}^{\mathrm{D},\,\mathrm{leximax}}(\vec{\varphi} + \varphi^{\mathtt{t}})] = \{\emptyset, a\}$ and $[\Delta_{\mu}^{\mathrm{D},\,\mathrm{leximax}}(\vec{\varphi} + \varphi^{\mathtt{f}})] = \{a\}$. If $n = 2k + 1$, take $[\varphi_1] = \cdots = [\varphi_k] = \{\emptyset\}$, $[\varphi_{k+1}] = \cdots = [\varphi_{2k}] = \{a\}$, $[\varphi_{2k+1}] = \{b\}$, $[\varphi^{\mathtt{t}}] = \{ab\}$ and $[\varphi^{\mathtt{f}}] = \{a\}$. We get that $[\Delta_{\mu}^{\mathrm{D},\,\mathrm{leximax}}(\vec{\varphi} + \varphi^{\mathtt{t}})] = \{\emptyset, a\}$ and $[\Delta_{\mu}^{\mathrm{D},\,\mathrm{leximax}}(\vec{\varphi} + \varphi^{\mathtt{f}})] = \{a\}$. Operators constructed with other aggregation functions are equivalent if the distance function is the drastic distance.

The story is different for the operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$: as seen in Proposition 5.12, constructive manipulation for skeptical acceptance, complete profiles, and $\mu = \top$ can usher an atom p into the result only if the agent does not believe p. In other words, the result can be changed with respect to p, but it is worth noting that the skeptical index does not increase by doing so. It turns out that this holds in general for the operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$ when the constraint is $\top$ i.e., this operator is strategyproof with respect to a dissatisfaction index i_{acc}, for $\mathrm{acc} \in \{\mathrm{skept}, \mathrm{cred}\}$.

Theorem 5.7

If all formulas in the profile, as well as $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$, are complete and $\mu = \top$, then the operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$ is strategyproof with respect to i_{skept} and i_{cred}.

Proof

For complete profiles and $\mu = \top$, the operator $\Delta^{\mathrm{H,\,sum}}$ returns models v that reflect majority opinion, i.e., if an atom p is true in a majority of formulas, p is in v; if p is false in a majority of formulas, then p is not in v; and if there is no majority (half of the formulas have p in their model), then the result contains both a v with p and a v' without p in them. A strategic agent cannot increase its index: adding something to its model can make this skeptically accepted, but this is not in the agent's belief. The reasoning for the other cases is similar.

The restrictions on $\Delta^{\mathrm{H,\,sum}}$ in Theorem 5.7 are essential: weakening any of them results in the operator being manipulable.

Proposition 5.13

If it is does not hold that $\mu = \top$ and all formulas in $\vec{\varphi}$, as well as the truthful position of the strategic agent, are complete, then $\Delta^{\mathrm{H,\,sum}}$ is manipulable with respect to i_{acc}.

Proof

We showcase, again, only instances of manipulation with respect to i_{skept} for the operator $\Delta^{\mathrm{H,\,sum}}$, as instances of manipulation with respect to i_{cred} are obtained by taking the duals. In the following we exhibit instances of manipulation in three cases, obtained by weakening the conditions of Theorem 5.7.

Case 1. Suppose $\mu = \top$ and every base except $\varphi^{\mathtt{t}}$ is required to be complete. Then we can find instances of manipulation for every profile of size $n \geq 1$. Take $[\varphi^{\mathtt{t}}] = \{a, b\}$.

For $n = 1$, take a profile $\vec{\varphi} = (\varphi_1)$, with $[\varphi_1] = \{a\}$. For $n \geq 2$ and $n = 2k$ take a profile $\vec{\varphi} = (\varphi_1, \ldots, \varphi_{2k})$ where $[\varphi_1] = \cdots = [\varphi_k] = \{\emptyset\}$ and $[\varphi_{k+1}] = \cdots = [\varphi_{2k}] = \{a\}$. For $n \geq 2$ and $n = 2k + 1$ take a profile $\vec{\varphi} = (\varphi_1, \ldots, \varphi_{2k})$ where $[\varphi_1] = \cdots = [\varphi_k] = \{\emptyset\}$ and $[\varphi_{k+1}] = \cdots = [\varphi_{2k+1}] = \{a\}$. In all cases, the profile $\vec{\varphi}$ is manipulable with respect to i_{skept} using $[\varphi^{\mathtt{f}}] = \{\emptyset\}$.

Case 2. Suppose, now, that $\varphi^{\mathtt{t}}$, $\varphi^{\mathtt{f}}$ and every base in $\vec{\varphi}$ is required to be complete, except one. Then we can still find instances of manipulation with respect to i_{skept}. For $n = 2$, take $\vec{\varphi} = (\varphi_1, \varphi_2)$, with $[\varphi_1] = \{a\}$, $[\varphi_1] = \{a, b\}$, $[\varphi^{\mathtt{t}}] = \{\emptyset\}$ and $[\varphi^{\mathtt{f}}] = \{b\}$.

Case 3. If every base in $\vec{\varphi}$ is complete, as well as $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$, but we are allowed to choose μ, then examples of manipulation are readily available. If $[\mu] = \{a, bc\}$, then we can take $\vec{\varphi} = (\varphi_1)$, with $[\varphi_1] = \{\emptyset\}$, $[\varphi^{\mathtt{t}}] = \{\emptyset\}$ and $[\varphi^{\mathtt{f}}] = \{b\}$. For a profile of size $n \geq 1$, taking $[\varphi_1] = \{a\}$ and $[\varphi_2] = \cdots = [\varphi_n] = \{\emptyset\}$, with $\varphi^{\mathtt{t}}$ and $\varphi^{\mathtt{f}}$ as before also results in an instance of manipulation with respect to i_{skept}.

Influence of one agent over the outcome

Up to now we have looked at whether the strategic agent can modify the merging result to its advantage. But it is useful to take a step back and ask whether the strategic agent can modify the result in the first place, i.e., whether it matters if the strategic agent takes part in the merging process at all and, if yes, how exactly it can influence it. Given a profile $\vec{\varphi}$, an operator Δ, a constraint μ and a formula φ, we say that $\Delta_\mu(\vec{\varphi})$ is *the intermediary result*, and $\Delta_\mu(\vec{\varphi}+\varphi)$ is *the final result*.

There are, *a priori*, two ways in which the agent can change the intermediary result: one is by removing interpretations from $[\Delta_\mu(\vec{\varphi})]$, i.e., by turning winning interpretations into non-winning interpretations; the other is by adding interpretations to $[\Delta_\mu(\vec{\varphi})]$, i.e., by turning non-winning interpretations into winners. If w is an interpretation, we say that the strategic agent *demotes* w *from* $\Delta_\mu(\vec{\varphi})$ *using* φ if $w \in [\Delta_\mu(\vec{\varphi})]$ and $w \notin [\Delta_\mu(\vec{\varphi}+\varphi)]$, and that it *promotes* w *with respect to* $\Delta_\mu(\vec{\varphi})$ *using* φ if $w \notin [\Delta_\mu(\vec{\varphi})]$ and $w \in [\Delta_\mu(\vec{\varphi}+\varphi)]$.

It turns out that for a significant proportion of the operators we are working with the strategic agent can demote any number of interpretations from the intermediary result, using an easy strategy: focus on the wanted interpretations, and submit a formula with those interpretations as models; the unwanted interpretations thus receive a penalty that renders them non-winning in the final result.

Proposition 5.14

If $\vec{\varphi}$ is a profile, μ is a constraint, $d \in \{d_{\mathrm{H}}, d_{\mathrm{D}}\}$, $\oplus \in \{\mathrm{sum}, \mathrm{leximax}, \mathrm{leximin}\}$ and $\mathcal{W} \subset [\Delta_\mu^{d,\oplus}(\vec{\varphi})]$ is a set of interpretations, then a strategic agent can demote all interpretations in $[\Delta_\mu^{d,\oplus}(\vec{\varphi})] \setminus \mathcal{W}$ from $\Delta_\mu^{d,\oplus}(\vec{\varphi})$ by submitting $\varphi_{\mathcal{W}}$ with $[\varphi_{\mathcal{W}}] = \mathcal{W}$.

Proof

The result can be shown using postulates M_{0-8}. It also follows directly using the resolvability postulate $\mathrm{M}_{\mathsf{RSVB}}$ from Section 5.3, which the operators considered here have been shown, in Proposition 5.9, to satisfy.

On the other hand, promoting interpretations is more difficult: the strategic agent's ability to promote an interpretation w depends on the margin by which w loses out to the winning interpretations. We will show this here for the operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$, after a detour through a couple of intermediate results. The first generalizes, in a way, the triangle inequality to the distances between a formula and two interpretations.

Lemma 5.4

If φ is a base and w_1 and w_2 are interpretations, then $d_{\mathrm{H}}(\varphi, w_1) \leq d_{\mathrm{H}}(\varphi, w_2) + d_{\mathrm{H}}(w_1, w_2)$.

	$\vec{\varphi}$	φ	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}+\varphi, \bullet)$
w_1	$\beta+\epsilon_1$	γ_1	$\beta+\gamma_1+\epsilon_1$
w_2	β	γ_2	$\beta+\gamma_2$

Table 5.9: Reversing the order between w_1 and w_2 by adding φ to $\vec{\varphi}$ is possible only if $\epsilon_1 \leq \delta_{1,2}$.

	$\vec{\varphi}$	$\{w_1\}$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}+\varphi, \bullet)$
w_1	$\beta+\epsilon_1$	0	$\beta+\epsilon_1$
w_2	β	$\delta_{1,2}$	$\beta+\delta_{1,2}$

Table 5.10: Reversing the order between w_1 and w_2 by adding φ to $\vec{\varphi}$, with $[\varphi]=\{w_1\}$, is possible if $\epsilon_1 \leq \delta_{1,2}$.

Proof

Suppose v_i is the model of φ at minimal Hamming distance to w_1 of all models of φ, i.e., $d_{\mathrm{H}}(\varphi, w_1) = d_{\mathrm{H}}(v_i, w_1)$, for $v_i \in [\varphi]$, and v_j is the model of φ at minimal Hamming distance to w_2 of all models of φ, i.e., $d_{\mathrm{H}}(\varphi, w_2) = d_{\mathrm{H}}(v_j, w_1)$, for $v_j \in [\varphi]$. By the regular triangle inequality we have that:

$$d_{\mathrm{H}}(w_1, v_j) \leq d_{\mathrm{H}}(w_1, w_2) + d_{\mathrm{H}}(w_2, v_j).$$

Since $d_{\mathrm{H}}(w_1, v_j) = d_{\mathrm{H}}(v_j, w_1)$, and $d_{\mathrm{H}}(v_i, w_1) \leq d_{\mathrm{H}}(v_j, w_1)$, the inequality just derived delivers the conclusion.

The following lemma uses this fact to determine what it takes for an interpretation to overtake another interpretation in the (H, sum)-induced preorder.

Lemma 5.5

If w_1 and w_2 are two interpretations such that $w_2 <_{\vec{\varphi}}^{\mathrm{H,\,sum}} w_1$, then there exists a base φ such that $w_1 \leq_{\vec{\varphi}+\varphi}^{\mathrm{H,\,sum}} w_2$ iff $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}, w_1) - d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}, w_2) \leq d_{\mathrm{H}}(w_1, w_2)$.

Proof

("⇒") We write $d_{\mathrm{H}}^{\mathrm{H,\,sum}}(\vec{\varphi}, w_2) = \beta$, $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}, w_1) = \beta+\epsilon_1$, with $\epsilon_1 > 0$. $d_{\mathrm{H}}(\varphi, w_1) = \gamma_1$ and $d_{\mathrm{H}}(\varphi, w_2) = \gamma_2$. This fits with the earlier naming convention, as w_2 is a winning interpretation in a direct contest with w_1 (i.e., if $[\mu] = \{w_1, w_2\}$). See Table 5.9 for a nicer picture of this situation. We have $w_1 \leq_{\vec{\varphi}+\varphi}^{\mathrm{H,\,sum}} w_2$ if and only if:

$$\beta + \gamma_1 + \epsilon_1 \leq \beta + \gamma_2. \tag{5.1}$$

By Lemma 5.4 we have that:

$$\gamma_2 \leq \gamma_1 + \delta_{1,2}. \tag{5.2}$$

Chaining inequalities 5.1 and 5.2 we obtain that $\beta+\gamma_1+\epsilon_1 \leq \beta+\gamma_1+\delta_{1,2}$. Simplifying, it follows that $\epsilon_1 \leq \delta_{1,2}$.

("⇐") Take φ such that $[\varphi] = \{w_1\}$. Then we have that $d_{\mathrm{H}}(\varphi, w_1) = 0$ and $d_{\mathrm{H}}(\varphi, w_1) = \delta_{1,2}$. This implies that $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi} + \varphi, w_1) = \beta + \epsilon_1$ and $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi} + \varphi, w_1) = \beta + \delta_{1,2}$. Since $\epsilon_1 \leq \delta_{1,2}$, it follows that $w_1 \leq_{\vec{\varphi}+\varphi}^{\mathrm{H,\,sum}} w_2$.

Intuitively, $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}, w) - d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}, w_i)$ is the margin by which w loses out to a winning interpretation w_i in $\leq_{\vec{\varphi}}^{d,\,\mathrm{sum}}$. Proposition 5.15 then tells us that the strategic agent can reverse the order between w and w_i if and only if this margin is less than the Hamming distance between w and w_i. In general, the amount of support the strategic agent can give to w relative to w_i is at most $d_{\mathrm{H}}(w, w_i)$ and thus, if w is trailing w_i by more than this amount, there is nothing the strategic agent can do for it. The main result follows now immediately.

Proposition 5.15

If w is an interpretation such that $w \in [\mu]$ and $w \notin [\Delta_{\mu}^{\mathrm{H,\,sum}}(\vec{\varphi})]$, then the strategic agent can promote w with respect to $\Delta_{\mu}^{\mathrm{H,\,sum}}(\vec{\varphi})$ iff $d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}, w) - d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}, w_i) \leq d_{\mathrm{H}}(w, w_i)$, for every $w_i \in [\Delta_{\mu}^{\mathrm{H,\,sum}}(\vec{\varphi})]$.

Proof

The claim follows from Lemma 5.5, as the agent has to reverse the order between w and every model of $\Delta_{\mu}^{\mathrm{H,\,sum}}(\vec{\varphi})$.

Using this result we also note that, if possible for an agent to promote an interpretation w, then it can do so using a complete formula.

Corollary 5.1

If the strategic agent can promote an interpretation w with respect to $\Delta_{\mu}^{d,\,\oplus}(\vec{\varphi})$, then it can do so with a formula φ_w such that $[\varphi_w] = \{w\}$.

This result is similar in spirit to Theorem 5.5, and suggests something like a best strategy if the goal is to promote w: the strategic agent can always submit a formula φ_w with w as the sole model, since if w can be promoted to the final result then φ_w is guaranteed to do it; otherwise, it does not matter what the agent submits.

	$\vec{\varphi}$	$\{w_1, w_2\}$	$\{w_4\}$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}+\varphi, \bullet)$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}+\varphi', \bullet)$
w_1	β	0	$\delta_{4,1}$	β	$\beta+\delta_{4,1}$
w_2	β	0	$\delta_{4,2}$	β	$\beta+\delta_{4,2}$
w_3	β	$\delta_{*,3}$	$\delta_{4,3}$	$\beta+\delta_{*,3}$	$\beta+\delta_{4,3}$
w_4	$\beta+\epsilon_4$	$\delta_{*,4}$	0	$\beta+\delta_{*,4}+\epsilon_4$	$\beta+\epsilon_4$

Table 5.11: The agent penalizes w_3 by not including it in the models of its reported formula, and can only promote w_4 if the margin ϵ_4 by which it trails the other interpretations is sufficiently small.

Example 5.13: Refining the intermediary result

Suppose $[\mu] = \{w_1, w_2, w_3, w_4\}$, $[\Delta_\mu^{d,\,\mathrm{sum}}(\vec{\varphi})] = \{w_1, w_2, w_3\}$, for $d \in \{d_{\mathrm{H}}, d_{\mathrm{D}}\}$. The strategic agent submits φ with $[\varphi] = \{w_1, w_2\}$. We write $d_{\mathrm{H}}(\vec{\varphi}, w_1) = d_{\mathrm{H}}(\vec{\varphi}, w_2) = d_{\mathrm{H}}(\vec{\varphi}, w_3) = \beta$, $d_{\mathrm{H}}(\vec{\varphi}, w_4) = \beta + \epsilon_4$ and $\delta_{*,3} = \min(\delta_{1,3}, \delta_{2,3})$, $\delta_{*,4} = \min(\delta_{1,4}, \delta_{2,4})$ for the distance from φ to w_3 and w_4, respectively. Table 5.11 offers an illustration. Notice now that $[\Delta_\mu^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi}+\varphi)] = \{w_1, w_2\}$, i.e., the strategic agent demotes w_3 from $\Delta_\mu^{d,\,\mathrm{sum}}(\vec{\varphi})$. To promote w_4 to the final result, the obvious strategy is for the strategic agent to submit φ', with $[\varphi'] = \{w_4\}$. In this case, promoting w_4 is successful only if $\epsilon_4 \leq \delta_{i,4}$, where $\delta_{i,4} = d_{\mathrm{H}}(w_i, w_4)$, for $i \in \{1, 2, 3\}$ (again, see Table 5.11). The same argument applies to the drastic distance d_{D}, the only difference being that $\delta_{*,3} = \delta_{*,4} = \delta_{i,4} = 1$, for $i \in \{1, 2, 3\}$.

5.5 Proportionality

In this section we study proportionality in the context of merging. Proportionality is one of the central fairness notions studied in social choice theory [Black, 1958, Dummett, 1984, Monroe, 1995], arising whenever a collective decision should reflect the amount of support in favor of a set of issues. Thus, notions of proportionality are key when it is desirable that preferences of larger groups have more influence on the outcome, while preferences of smaller groups are not neglected. The idea of proportional representation shows up in many application scenarios: it is a key ingredient of parliamentary elections [Balinski and Young, 1982] and, more generally, of multiwinner voting, i.e., the task of electing a committee of multiple candidates [Faliszewski et al., 2017a], of which we have already had a taste in Section 2.4 in the context of ABC social choice functions. That proportionality issues are relevant to merging can be readily illustrated using our old friends, the Academy members trying to come up with a list of nominees for the category of *Best Director*.

d_{H}	$[\varphi_i]$, for $1 \leq i \leq 4$ $4 \cdot a_1a_2a_3a_4a_5$	$[\varphi_5]$ $b_1b_2b_3b_4b_5$	$d_{\mathrm{H}}^{\mathrm{sum}}(\vec{\varphi}, \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximax}}(\vec{\varphi}, \bullet)$	$d_{\mathrm{H}}^{\mathrm{leximin}}(\vec{\varphi}, \bullet)$
$a_1a_2a_3a_4a_5$	$4 \cdot 0$	10	**10**	$(10, 0, 0, 0, 0)$	$(\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{10})$
$a_1a_2a_3a_4b_1$	$4 \cdot 2$	8	16	$(8, 2, 2, 2, 2)$	$(2, 2, 2, 2, 8)$
$a_1a_2a_3b_1b_2$	$4 \cdot 4$	6	22	$(\mathbf{6}, \mathbf{4}, \mathbf{4}, \mathbf{4}, \mathbf{4})$	$(4, 4, 4, 4, 6)$
$a_1a_2b_1b_2b_3$	$4 \cdot 6$	4	28	$(6, 6, 6, 6, 4)$	$(4, 6, 6, 6, 6)$
$a_1b_1b_2b_3b_4$	$4 \cdot 8$	2	34	$(8, 8, 8, 8, 2)$	$(2, 8, 8, 8, 8)$
$b_1b_2b_3b_4b_5$	$4 \cdot 10$	0	40	$(10, 10, 10, 10, 0)$	$(0, 10, 10, 10, 10)$
...					

Table 5.12: Hamming distances from each formula in the profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq 5}$ to a representative sample of intepretations of size 5, together with the aggregated distances according to the sum, leximax and leximin aggregation functions. The optimal outcomes are those that minimize overall distance. None of these methods picks out the proportional outcomes, e.g., $a_1a_2a_3a_4b_1$.

Example 5.14: #OscarsSoUnrepresentative

To make the proportionality issues more apparent, we slightly alter the scenario of Example 1.5: suppose the list of nominees has to contain five directors, and there are now ten names being circulated. There are five Academy members whose opinions are divided along two distinct and opposing camps: the first four members support five of the names, while the remaining member supports the other four names.

We model this as a merging task where the set of atoms is $A \cup B$, with $A = \{a_1, a_2, a_3, a_4, a_5\}$ and $B = \{b_1, b_2, b_3, b_4, b_5\}$. The profile is $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq 5}$, with $[\varphi_1] = [\varphi_2] = [\varphi_3] = [\varphi_4] = \{a_1a_2a_3a_4a_5\}$ and $[\varphi_5] = \{b_1b_2b_3b_4b_5\}$. The constraint μ is a propositional formula such that $[\mu] = \{a_1a_2a_3a_4a_5, a_1a_2a_3a_4b_1, \dots\}$, i.e., a propositional formula that encodes the cardinality requirement on the output and whose models make exactly five atoms true. Both the formula μ and its set of models are too large to write here in full, but an illustration of the relevant scores for the main operators is offered in Table 5.12.

We obtain that $[\Delta_{\mu}^{\mathrm{H},\,\mathrm{sum}}(\vec{\varphi})] = \{a_1a_2a_3a_4a_5\}$, $[\Delta_{\mu}^{\mathrm{H},\,\mathrm{leximax}}(\vec{\varphi})] = \{a_1a_2a_3b_1b_2\}$ and $[\Delta_{\mu}^{\mathrm{H},\,\mathrm{leximin}}(\vec{\varphi})] = \{a_1a_2a_3a_4a_5\}$. Note that the operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$ selects the outcome that is supported by the majority of the members of the profile, while the operator $\Delta^{\mathrm{H},\,\mathrm{leximax}}$ attempts to steer a middle ground between the two groups, selecting an outcome that has, roughly, the same number of names from each camp. The operator $\Delta^{\mathrm{H},\,\mathrm{leximin}}$ tries to improve the situation of the best off agent, which in this case corresponds to the result of operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$. However, it can be argued that a proportional outcome would reflect the composition of the profile and select four atoms from A and one from B.

The situation presented in Example 5.14 dovetails with the observation made at the

end of Section 3.4, according to which the operator $\Delta^{\mathrm{H},\,\mathrm{sum}}$ has majoritarian tendencies, $\Delta^{\mathrm{H},\,\mathrm{leximax}}$ is egalitarian and $\Delta^{\mathrm{H},\,\mathrm{leximin}}$ is elitist. Proportional outcomes, as seen in Example 5.14, can fall in between these extremes, and the existing merging operators can fail to pick them out. Our aim is to find a compromise between the majoritarian, egalitarian and elitist merging operators by formalizing proportionality postulates and proposing concrete merging operators that deliver proportional results.

Satisfaction-based merging operators

In defining proportional belief merging operators we rely on the Proportional Approval Voting (PAV) rule for multiwinner elections, presented in Section 2.4 and known to satisfy particularly strong proportionality requirements [Aziz et al., 2017]. Since the PAV rules is based on maximizing overall satisfaction, we introduce an alternative way of representing merging operators, based on satisfaction. The key to doing so is a *satisfaction measure* s, which is a function $s\colon \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ quantifying the amount of satisfaction $s(v, w)$ of interpretation v with interpretation w. If v and w are interpretations, we write $s(v, w)$ for the satisfaction of v with w. Using a satisfaction measure s, we build our way towards full-fledged merging operators in the familiar manner. We first lift the satisfaction notion to the *satisfaction* $s(\varphi, w)$ *of a formula* φ *with* w, defined as:

$$s(\varphi, w) = \max(s(v, w))_{v \in [\varphi]}.$$

The second ingredient is an aggregation function $\oplus$ used to obtain the collective satisfaction of a profile $\vec{\varphi}$ with an interpretation w. In this section we will use the sum aggregation function exclusively, so it becomes unnecessary to make $\oplus$ an explicit parameter in the notation. Thus, if $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ is an $\mathcal{L}$-profile and w is an interpretation, the *collective satisfaction* $s(\vec{\varphi}, w)$ *of a profile* $\vec{\varphi}$ *with* w is defined as

$$s^{\mathrm{sum}}(\vec{\varphi}, w) = \sum_{\varphi \in \vec{\varphi}} s(\varphi, w).$$

As already mentioned, we will write simply $s(\vec{\varphi}, w)$ instead of $s^{\mathrm{sum}}(\vec{\varphi}, w)$. Using the satisfaction indices we can order interpretations according to their satisfaction with respect to φ and $\vec{\varphi}$. Thus, the *s-induced rankings* $\geq^s_\varphi$ and $\geq^s_{\vec{\varphi}}$ are defined, respectively, as:

$$w_1 \geq^s_\varphi w_2 \text{ if } s(\varphi, w_1) \geq s(\varphi, w_2),$$
$$w_1 \geq^s_{\vec{\varphi}} w_2 \text{ if } s(\vec{\varphi}, w_1) \geq s(\vec{\varphi}, w_2).$$

Finally, if s is a satisfaction measure, the *s-induced merging operator* Δ^s is defined, for any $\mathcal{L}$-profile $\vec{\varphi} = (\varphi_i)_{1 \leq i \leq n}$ and propositional formula μ, as a propositional formula $\Delta^s_\mu(\vec{\varphi})$ such that:

$$[\Delta^s_\mu(\vec{\varphi})] \stackrel{\mathrm{def}}{=} \max_{\geq^s_{\vec{\varphi}}}[\mu].$$

In other words, $\Delta^s_\mu(\vec{\varphi})$ is a formula whose models are exactly the models of μ that maximize overall satisfaction of $\vec{\varphi}$.

Note that we can convert a distance-based merging operator $\Delta^{d,\,\mathrm{sum}}$ into an equivalent satisfaction-based operator by inverting the pseudo distance d, i.e., by defining a satisfaction measure s as $s(v,w) = m - d(v,w)$, for any interpretations v and w (recall that m is the number of atoms in A). The resulting satisfaction-based operator is such that $\Delta^{s}_{\mu}(\vec{\varphi}) \equiv \Delta^{d,\,\mathrm{sum}}\mu(\vec{\varphi})$, for any profile $\vec{\varphi}$ and μ. Note, also, that since d is a distance and thus symmetric (i.e., $d(v,w) = d(w,v)$, for any interpretations v and w), the satisfaction measure s defined on the basis of it is also symmetric. This being said, we do not require satisfaction measures to be symmetric in general. Consequently, satisfaction-based operators as defined here form a more general class than distance-based operators $\Delta^{d,\,\mathrm{sum}}$, where d is a distance. It is worth mentioning that the satisfaction-based approach we propose here is not a mere stylistic variant of the distance-based view; it also encourages a different viewpoint on merging, where the goal is to find an outcome making agents happy, subject to fairness norms. Scenarios where this viewpoint is warranted occur most of all in social choice settings, and a key strength of merging is its ability to accommodate them.

The concrete satisfaction measures we propose are defined, for any interpretations v and w, as follows:

$$\begin{array}{lrcl}
\text{approval-based}\left\{\begin{array}{r} \\ \\ \end{array}\right. & s_{\mathrm{AV}}(v,w) & \stackrel{\mathrm{def}}{=} & |v \cap w|, \\
& s_{\mathrm{PAV}}(v,w) & \stackrel{\mathrm{def}}{=} & h(|v \cap w|), \\
& s_{\mathrm{bPAV}}(v,w) & \stackrel{\mathrm{def}}{=} & 2h(|v \cap w|) - h(|w|), \\
\text{binary sat.-based}\left\{\begin{array}{r} \\ \\ \end{array}\right. & s_{\mathrm{hH}}(v,w) & \stackrel{\mathrm{def}}{=} & h(m - d_{\mathrm{H}}(v,w)), \\
& s_{\mathrm{hD}}(v,w) & \stackrel{\mathrm{def}}{=} & h(m - d_{\mathrm{D}}(v,w)).
\end{array}$$

The satisfaction measures are divided into two groups and, predictably, generate two groups of operators. The *approval-based* operators, consisting of the AV *operator* Δ^{AV}, the PAV *operator* Δ^{PAV} and the *bounded* PAV *operator* Δ^{bPAV}, mimic the behavior of an ABC voting rule, as described in Section 2.4, in that they compute satisfaction of v with w based on how many atoms v and w have in common, similarly to how satisfaction of an approval ballot A_i with a potential committee w is based on how many approved candidates in A_i find themselves in w. Note that, while an ABC voting rule is defined only for committees of fixed size, the merging operators we propose select among interpretations of any size. Nonetheless, it is straightforward to see that if the allowed outcomes (here, models of the constraint μ) are restricted to a given size, then the operators Δ^{PAV} and Δ^{bPAV} are equivalent and extend the PAV voting rule.

The operator Δ^{AV} is put forward as a benchmark approval-based operator, based on a satisfaction measure that simply counts the atoms v and w have in common: in particular, Δ^{AV} does not incorporate any proportionality ideas. Consequently, as shown in Proposition 5.16, the Δ^{AV} operator does not extend PAV and does not meet any of the proportionality requirements we expect.

	s_{PAV}			s_{bPAV}		
	a_1	a_1a_2	sum	a_1	a_1a_2	sum
a_1	$h(1)$	$h(1)$	2	$2h(1)-h(1)$	$2h(1)-h(1)$	2
a_1a_2	$h(1)$	$h(2)$	2.5	$2h(1)-h(2)$	$2h(2)-h(2)$	2

Table 5.13: Behavior of Δ^{PAV} and Δ^{bPAV} regarding M_4. The operator Δ^{PAV} does not satisfy postulate M_4 for the profile $\vec{\varphi}$ and constraint μ, though the operator Δ^{bPAV} does.

Proposition 5.16

The approval-based merging operator Δ^{AV} does not extend PAV.

Proof

For the set of atoms $A \cup B$, $\vec{\varphi}$ and μ as in Example 5.14, it holds that $[\Delta^{\text{AV}}_{\mu}(\vec{\varphi})] = \{a_1a_2a_3a_4a_5\}$, whereas the PAV outcome in the corresponding ABC election selects interpretations that make exactly four atoms from A and one from B true.

The Δ^{PAV} operator refines Δ^{AV} by using the harmonic function h, which is known to behave well with respect to proportionality requirements [Aziz et al., 2017]. Intuitively, the harmonic function reflects the "diminishing returns" of added satisfaction: the difference between $h(x)$ and $h(x+1)$ gets smaller as x increases. Thus, the operator Δ^{PAV} is a prime candidate for a proportional satisfaction-based merging operator. Nonetheless, Δ^{PAV} has several shortcomings, which serve as motivation for the remaining operators.

One drawback of Δ^{PAV} is that it favors larger interpretations if available, as shown in Example 5.15, i.e., it tries to increase agents' satisfaction by setting as many atoms to true as possible. Such an inflationary strategy may be undesirable in practice and, in a belief merging setting, interferes with postulate M_4.

Example 5.15: Δ^{PAV} does not satisfy postulate M_4

For the set of atoms $A = \{a_1, a_2\}$, profile $\vec{\varphi} = (\varphi_1, \varphi_2)$, with $[\varphi_1] = \{a_1\}$ and $[\varphi_2] = \{a_1a_2\}$, and constraint μ such that $[\mu] = \{a_1, a_1a_2\}$, we obtain that $[\Delta^{\text{PAV}}_{\mu}(\vec{\varphi})] = \{a_1a_2\}$, contradicting postulate M_4. The same result is obtained for Δ^{AV}, but $[\Delta^{\text{bPAV}}_{\mu}(\vec{\varphi})] = \{a_1, a_1a_2\}$, which is in accordance with M_4. The situation is depicted in Table 5.13.

To curb the inflationary tendencies of Δ^{PAV}, operator Δ^{bPAV} introduces a penalty on interpretations depending on their size, in the process ensuring satisfaction of postulate M_4 as well. Indeed, Δ^{bPAV} is recommended by the fact, which we will elaborate on shortly, that it is the *only* operator from a fairly broad class that manages to balance

proportionality and fairness as formalized by postulate M_4. Note, however, that s_{bPAV} is not symmetric.

Example 5.16: s_{bPAV} is not symmetric

It holds that $s_{\text{bPAV}}(a, ab) < s_{\text{bPAV}}(ab, a)$. Intuitively, this means it is worse to obtain b if it is not wanted than to not obtain it if it is wanted.

A related problem with Δ^{PAV} stems from the fact that $s_{\text{PAV}}(v, w)$ is obtained by counting only atoms v and w have in common. Hence, Δ^{PAV} has a bias towards positive literals, and is insensitive to the presence of extraneous, possibly unwanted atoms in w, the assumption being that atoms in w that are not in v represent issues on which v has no opinion on, and thus their presence has no effect on the satisfaction of v (see Example 5.17). This assumption is not always justified, and it turns out to interfere with postulate M_2.

The *binary satisfaction-based* operators, consisting of the *harmonic drastic operator* Δ^{hD} and the *harmonic Hamming operator* Δ^{hH}, are introduced in an attempt to deal with the effect of unwanted atoms while, at the same time, providing proportional outcomes. The satisfaction measures they are based on penalize interpretations w if they include atoms for which an explicit preference is not stated. This is done by inverting familiar notions of distance between interpretations, which pay attention to atoms appearing in one of the interpretation but not in the other, and leads to an equal treatment of positive and negative literals. The harmonic function h is applied to this satisfaction notion, with the idea of ensuring proportionality. The operators that emerge are worth investigating: neither of them extends PAV as hinted at in Example 5.17, but from this point onward their properties diverge. Though Δ^{hH} does not extend PAV, it still ends up having interesting proportionality properties, formalized shortly.

Example 5.17: Approval-based operators and postulate M_2

For interpretations a_1 and a_1a_2, it holds that $s_{\text{PAV}}(a_1, a_1) = s_{\text{PAV}}(a_1, a_1a_2)$, i.e., according to the PAV satisfaction index, a_1 and a_1a_2 are indistinguishable for a_1. The assumption behind s_{PAV} is that an agent who wants a_1 is equally satisfied with a_1a_2 as it is with a_1, i.e., is not bothered by the presence of a_2. This attitude results in postulate M_2 not being satisfied.

For $A = \{a_1, a_2\}$, $\vec{\varphi} = (\varphi)$, $[\varphi] = \{a_1\}$ and $\mu = \top$, we obtain that $[\Delta_\mu^{\text{PAV}}(\vec{\varphi})] = \{a_1, a_1a_2\}$, contrary to M_2. Whereas satisfaction of postulate M_2 would require the result to be $\{a_1\}$. On the other hand, $s_{\text{hH}}(a_1, a_1) = h(2)$, while $s_{\text{hH}}(a_1, a_1a_2) = h(1)$ and $[\Delta_\mu^{\text{hH}}(\vec{\varphi})] = \{a_1\}$, in accordance with M_2. Thus, according to s_{hH}, a_1a_2 provides less satisfaction to a_1 than a_1 alone, i.e., the agent is bothered by the presence of a_2, for which an explicit preference was not stated. Presumably, this is because the presence of a_2 is unwanted and contributes negatively towards the final amount of satisfaction. Consequently, for $\vec{\varphi}$ and μ as above, $[\Delta_\mu^{\text{hH}}(\vec{\varphi})] = \{a_1\}$. This is in accord

with postulate M_2.

The operator Δ^{hD} turns out to be so coarse in its assessment of satisfaction as to become, as we now show, indistinguishable from existing merging operators defined using drastic distance d_{D}. We arrive at this via some intermediary notions and results, the first of which is a satisfaction measure that inverts the drastic distance d_{D}. Thus, the satisfaction measure s_{sD} is defined as $s_{\text{sD}}(v, w) \stackrel{\text{def}}{=} m - d_{\text{D}}(v, w)$, for any interpretations v and w. The sD-induced merging operator defined using the satisfaction measure s_{sD} is denoted as Δ^{sD}. The first thing we show is that Δ^{hD} and Δ^{sD} are equivalent.

Lemma 5.6

The satisfaction-based operators Δ^{hD} and Δ^{sD} are equivalent.

Proof

Take a profile $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$ and two interpretations $w_1, w_2 \in [\mu]$. We will show that $s_{\text{sD}}(\vec{\varphi}, w_1) \geq s_{\text{sD}}(\vec{\varphi}, w_2)$ if and only if $s_{\text{hD}}(\vec{\varphi}, w_1) \geq s_{\text{hD}}(\vec{\varphi}, w_2)$.

We will denote by a_i the number of formulas φ in $\vec{\varphi}$ such that $w_i \in [\varphi]$, and by b_i the number of formulas φ in $\vec{\varphi}$ such that $w_i \notin [\varphi]$, for $i \in \{1, 2\}$. It then holds that $a_1 + b_1 = a_2 + b_2 = n$ and that:

$$\begin{aligned} s_{\text{sD}}(\vec{\varphi}, w_i) &= a_i m + b_i(m-1), \\ s_{\text{hD}}(\vec{\varphi}, w_i) &= a_i h(m) + b_i h(m-1), \end{aligned}$$

for $i \in \{1, 2\}$. The claim we want to prove translates as:

$$\begin{aligned} a_1 m + b_1(m-1) \geq a_2 m + b_2(m-1) \qquad & \text{iff} \\ a_1 h(m) + b_1 h(m-1) \geq a_2 h(m) + b_2 h(m-1). & \end{aligned}$$

With some algebraic manipulation of the left-hand-side term, and using the fact that $a_1 + b_1 = a_2 + b_2 = n$, we obtain that:

$$\begin{aligned} a_1 m + b_1(m-1) &\geq a_2 m + b_2(m-1) && \text{iff} \\ (a_1 + b_1)m - b_1 &\geq (a_2 + b_2)m - b_2 && \text{iff} \\ nm - b_1 &\geq nm - b_2 && \text{iff} \\ b_2 &\geq b_1 && \text{iff} \\ n - a_2 &\geq n - a_1 && \text{iff} \\ a_1 &\geq a_2. \end{aligned}$$

With some algebraic manipulation of the right-hand-side term, and using the facts that $h(m) = h(m-1) + \frac{1}{m}$ and $a_1 + b_1 = a_2 + b_2 = n$, we obtain that:

$$\begin{aligned}
a_1 h(m) + b_1 h(m-1) &\geq a_2 h(m) + b_2 h(m-1) && \text{iff} \\
a_1(h(m-1) + \frac{1}{m}) + b_1 h(m-1) &\geq a_2(h(m-1) + \frac{1}{m}) + b_2 h(m-1) && \text{iff} \\
(a_1 + b_1)h(m-1) + a_1(\frac{1}{m}) &\geq (a_2 + b_2)h(m-1) + a_2(\frac{1}{m}) && \text{iff} \\
nh(m-1) + a_1(\frac{1}{m}) &\geq nh(m-1) + a_2(\frac{1}{m}) && \text{iff} \\
a_1 &\geq a_2.
\end{aligned}$$

Thus, both sides reduce to the same inequality, and are therefore equivalent. Moreover, it is straightforward to see that equality is obtained on both sides in the same case: when there are as many formulas in $\vec{\varphi}$ that feature w_1 as a model as there are formulas that feature w_2 as a model. In other words, we have that:

$$\begin{aligned}
s_{\text{sD}}(\vec{\varphi}, w_1) = s_{\text{sD}}(\vec{\varphi}, w_2) \text{ iff } s_{\text{hD}}(\vec{\varphi}, w_1) &= s_{\text{hD}}(\vec{\varphi}, w_2) \\
\text{iff } a_1 &= a_2.
\end{aligned}$$

We have obtained, therefore, that $s_{\text{sD}}(\vec{\varphi}, w_1) \geq s_{\text{sD}}(\vec{\varphi}, w_2)$ if and only if $s_{\text{hD}}(\vec{\varphi}, w_1) \geq s_{\text{hD}}(\vec{\varphi}, w_2)$. This, now, implies the conclusion, namely that $\Delta_\mu^{\text{hD}}(\vec{\varphi}) \equiv \Delta_\mu^{\text{sD}}(\vec{\varphi})$, for any constraint μ.

We can now piece the details of these results together and conclude that the operator Δ^{hD} is identical to the existing merging operator Δ^{D}.

Theorem 5.8

The satisfaction-based operator Δ^{hD} is equivalent to the distance-based operator Δ^{D}.

Proof

By Lemma 5.6, operator Δ^{hD} is equivalent to Δ^{sD} defined previously. It is now straightforward to see that Δ^{sD} is equivalent to Δ^{D}.

As a result, the Δ^{hD} operator is not responsive to proportionality requirements. What emerges, therefore, is a landscape with three merging operators relevant to the issue of proportionality, i.e., Δ^{PAV}, Δ^{bPAV} and Δ^{hH}. Out of these, Δ^{bPAV} and Δ^{hH} address, each in its own way, problems arising with the Δ^{PAV} operator: Δ^{bPAV} penalizes interpretations according to their size, while Δ^{hH} uses an approach reminiscent from logic, where positive and negative literals are treated equally. As we will see, the proposed solutions involve various trade-offs between proportionality and postulates M_{0-8}. The first result in that

direction shows that any satisfaction-based operator satisfies a core set of these postulates.

Proposition 5.17

If s is a satisfaction measure, then the merging operator Δ^s satisfies postulates M_{0-1}, M_3 and M_{5-8}.

Proof

Using the definition of the satisfaction-based operator Δ^s we infer that $\emptyset \subset [\Delta^s_\mu(\vec{\varphi})] \subseteq [\mu]$, i.e., Δ^s is a formula whose set of models is a non-empty subset of the set of models of μ, which implies that postulates M_{0-1} are satisfied. Since $\Delta^s_\mu(\vec{\varphi})$ is defined solely in terms of its models, the syntax of the formulas involved does not influence the merging result and, hence, postulate M_3 is satisfied.

For postulate M_5, take an interpretation $w \in [\Delta^s_\mu(\vec{\varphi}_1) \wedge \Delta^s_\mu(\vec{\varphi}_2)]$, and an arbitrary interpretation $w' \in [\mu]$. We have that:

$$\sum_{\varphi \in \vec{\varphi}_1} s(\varphi, w) \geq \sum_{\varphi \in \vec{\varphi}_1} s(\varphi, w'), \tag{5.3}$$

$$\sum_{\varphi \in \vec{\varphi}_2} s(\varphi, w) \geq \sum_{\varphi \in \vec{\varphi}_2} s(\varphi, w'). \tag{5.4}$$

Adding the two inequalities gives us:

$$\sum_{\varphi \in \vec{\varphi}_1} s(\varphi, w) + \sum_{\varphi \in \vec{\varphi}_2} s(\varphi, w) \geq \sum_{\varphi \in \vec{\varphi}_1} s(\varphi, w') + \sum_{\varphi \in \vec{\varphi}_2} s(\varphi, w'),$$

which, in turn, implies that:

$$\sum_{\varphi \in (\vec{\varphi}_1 + \vec{\varphi}_2)} s(\varphi, w) \geq \sum_{\varphi \in (\vec{\varphi}_1 + \vec{\varphi}_2)} s(\varphi, w'). \tag{5.5}$$

Thus, the interpretation w, which provides maximal satisfaction for profiles $\vec{\varphi}_1$ and $\vec{\varphi}_2$, also provides maximal satisfaction for profile $\vec{\varphi}_1 + \vec{\varphi}_2$, which allows us to conclude that $w \in [\Delta^s_\mu(\vec{\varphi}_1 + \vec{\varphi}_2)]$.

For postulate M_6 notice that if one of the inequalities 5.3 or 5.4 is strict, then inequality 5.5 is also strict. Thus, if interpretation w' does not provide maximal satisfaction with respect to $\vec{\varphi}_1$ or $\vec{\varphi}_2$, then it does not provide maximal satisfaction with respect to $\vec{\varphi}_1 + \vec{\varphi}_2$ either. In other words, if $w' \notin [\Delta^s_\mu(\vec{\varphi}_1) \wedge \Delta^s_\mu(\vec{\varphi}_2)]$, then $w' \notin [\Delta^s_\mu(\vec{\varphi}_1 + \vec{\varphi}_2)]$, which proves the claim.

For postulate M_7, we have that if $w \in [\Delta^s_{\mu_1}(\vec{\varphi}) \wedge \mu_2]$, then $s(\vec{\varphi}, w) \geq s(\vec{\varphi}, w')$, for any $w' \in [\mu_1]$. Since $[\mu_1 \wedge \mu_2] \subseteq [\mu_1]$, it is straightforward to conclude from here that $s(\vec{\varphi}, w) \geq s(\vec{\varphi}, w')$, for any $w' \in [\mu_1 \wedge \mu_2]$, i.e., if w provides maximal satisfaction when

the available options are the models of μ_1, it will also provide maximal satisfaction when we restrict the available options to the models of $\mu_1 \wedge \mu_2$. Since $w \in [\mu_2]$ as well, it follows that $w \in [\Delta^s_{\mu_1 \wedge \mu_2}(\vec{\varphi})]$.

Conversely, for postulate M_8, suppose $w \in [\Delta^s_{\mu_1 \wedge \mu_2}(\vec{\varphi})]$ and suppose $w \notin [\Delta^s_{\mu_1}(\vec{\varphi}) \wedge \mu_2]$. This means that $w \notin [\Delta^s_{\mu_1}(\vec{\varphi})]$. Since $\Delta^s_{\mu_1}(\vec{\varphi}) \wedge \mu_2$ is consistent, there exists $w' \in [\Delta^s_{\mu_1}(\vec{\varphi}) \wedge \mu_2]$, which, together with the finding that $w \notin [\Delta^s_{\mu_1}(\vec{\varphi})]$, implies that $s(\vec{\varphi}, w') > s(\vec{\varphi}, w)$. However, from the assumption that $w \in [\Delta^s_{\mu_1 \wedge \mu_2}(\vec{\varphi})]$ we obtain that $s(\vec{\varphi}, w) \geq s(\vec{\varphi}, w')$, which leads to a contradiction.

Proposition 5.17 applies to both the approval-based and the harmonic distance-based operators. What remains, then, is an understanding of how the new satisfaction measures interact with postulates M_2 and M_4, and we settle the issue by characterizing the types of satisfaction measures compliant with these postulates. If v and w are interpretations such that $v \neq w$, the following properties prove to be relevant:

(S_1) $s(v, v) > s(v, w)$;

(S_2) $s(v, v) > s(w, v)$;

(S_3) $s(v, v) = s(w, w)$;

(S_4) $s(v, w) = s(w, v)$.

Properties S_{1-4} formalize the intuition that satisfaction is symmetric (S_4), maximal when one obtains *exactly* what one wants, and trailing off as the outcome diverges from one's most desired outcome (S_{1-3}). Theorem 5.9 shows that properties S_{1-3} capture satisfaction measures compliant with postulate M_2.

Theorem 5.9

A satisfaction-based merging operator Δ^s satisfies postulate M_2 if and only if s satisfies properties S_{1-3}.

Proof

("⇒") Take a satisfaction-based merging operator Δ^s that satisfies postulate M_2. We will show that s satisfies property S_{1-3}.

For property S_1, take interpretations v and w such that $v \neq w$. Consider, now, formulas φ and μ such that $[\varphi] = \{v\}$ and $[\mu] = \{v, w\}$, and the profile $\vec{\varphi} = (\varphi)$. applying postulate M_2, we have that $[\Delta^s_\mu(\vec{\varphi})] = [\varphi \wedge \mu] = \{v\}$. This implies that $v \in \max_{\geq^s_{\vec{\varphi}}}[\mu]$ and $w \notin \max_{\geq^s_{\vec{\varphi}}}[\mu]$, which leads to $s(\varphi, v) > s(\varphi, w)$. This, in turn, implies that $s(v, v) > s(v, w)$.

	v	w	max
v	$s(v,v)$	$s(w,v)$	$\max\{s(v,v), s(w,v)\}$
w	$s(v,w)$	$s(w,w)$	$\max\{s(v,w), s(w,w)\}$

Table 5.14: Satisfaction indices when $\vec{\varphi} = (\varphi)$, $[\varphi] = [\mu] = \{v, w\}$. The models of φ are written on the top row; the columns indicate models of μ.

	$[\varphi_1]$ $\{v_1, v_2, \ldots\}$	$\ldots$ $\ldots$	$[\varphi_n]$ $\{v_1, v_2, \ldots\}$	$\sum$
v_1	$s(v_1, v_1)$	$\ldots$	$s(v_1, v_1)$	$ns(v_1, v_1)$
v_2	$s(v_2, v_2)$	$\ldots$	$s(v_2, v_2)$	$ns(v_2, v_2)$
w	$\max_{v \in [\varphi_1]} s(v, w)$	$\ldots$	$\max_{v \in [\varphi_n]} s(v, w)$	$\sum_{i=1}^{n} \max_{v \in [\varphi_i]} s(v, w)$

Table 5.15: Satisfaction indices when $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$, $v_1, v_2 \in [(\bigwedge_{\varphi_i \in \vec{\varphi}} \varphi_i) \wedge \mu]$ and $w \in [\mu]$ but $w \notin [\bigwedge_{\varphi_i \in \vec{\varphi}}]$.

For property S_2, suppose there exist interpretations v and w such that $v \neq w$ and $s(v, v) \leq s(w, v)$. Take, now, a formula φ such that $[\varphi] = \{v, w\}$ and μ as before, with $[\mu] = \{v, w\}$ (see Table 5.14). Our assumptions, together with property S_1, proven above, allow us to conclude that:

$$s(v, w) < s(v, v) \leq s(w, v) < s(w, w).$$

In other words, $\max\{s(v, v), s(w, v)\} = s(w, v)$ and $\max\{s(v, w), s(w, w)\} = s(w, w)$, which means that:

$$\max\{s(v, v), s(w, v)\} < \max\{s(v, w), s(w, w)\}.$$

But, by postulate M_2, we have that $[\Delta^s_\mu(\vec{\varphi})] = \{v, w\}$ and thus it holds that $\max\{s(v, v), s(w, v)\} = \max\{s(v, w), s(w, w)\}$, which leads to a contradiction, and to the conclusion that property S_2 holds.

Finally, taking φ and μ as in the proof for property S_2, and using the result derived there, we conclude that $\max\{s(v, v), s(w, v)\} = s(v, v)$ and that $\max\{s(v, w), s(w, w)\} = s(w, w)$. Postulate M_2, now, implies that $s(v, v) = s(w, w)$ and hence property S_3 is satisfied.

("$\Leftarrow$") Conversely, we want to show that if s satisfies properties S_{1-3}, then Δ^s satisfies postulate M_2. To that end, take a profile $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$ and a formula μ such that $(\bigwedge_{\varphi_i \in \vec{\varphi}} \varphi_i) \wedge \mu$ is consistent. We will prove the claim in two steps. First, we show that for any interpretations $v_1, v_2 \in [(\bigwedge_{\varphi_i \in \vec{\varphi}} \varphi_i) \wedge \mu]$, we have that $s(\vec{\varphi}, v_1) = s(\vec{\varphi}, v_2)$. Then, we show, that if w is an interpretation such that $w \in [\mu]$ but $w \notin [\bigwedge_{\varphi_i \in \vec{\varphi}} \varphi_i]$, then $s(\vec{\varphi}, w) < s(\vec{\varphi}, v_1) = s(\vec{\varphi}, v_2)$.

Indeed, if $v_1 = v_2$, then the first claim is immediate. If $v_1 \neq v_2$, then we reason as follows. Take a formula $\varphi_i \in \vec{\varphi}$. Using the fact that $v_1 \in [\varphi_i]$ and property S_2, we get that $s(v_1, v_1) > s(v_j, v_1)$, for any $v_j \in [\varphi_i]$ such that $v_j \neq v_1$. Thus, $s(\varphi_i, v_1) = s(v_1, v_1)$, for any $\varphi_i \in \vec{\varphi}$, and it follows that $s(\vec{\varphi}, v_1) = ns(v_1, v_1)$ (see Table 5.15). Analogously, we get that $s(\varphi_i, v_2) = s(v_2, v_2)$, for any $\varphi_i \in \vec{\varphi}$ and $s(\vec{\varphi}, v_1) = ns(v_2, v_2)$. By property S_3, we have that $s(v_1, v_1) = s(v_2, v_2)$. This, in turn, implies that $s(\vec{\varphi}, v_1) = s(\vec{\varphi}, v_2)$.

For the second claim, we have that $s(\varphi_i, w) = \max_{v \in [\varphi_i]} s(v, w)$, for any $\varphi_i \in \vec{\varphi}$. By property S_1, we have that $\max_{v \in [\varphi_i]} s(v, w) \leq s(v_1, v_1)$. Equality is achieved if $w \in [\varphi_i]$: however, we have assumed that $w \notin [\bigwedge_{\varphi_i \in \vec{\varphi}}]$, and thus there exists at least one $\varphi_i \in \vec{\varphi}$ such that $w \notin [\varphi_i]$. In other words, at least one of the inequalities is strict. Hence, when we add up all the satisfaction indices for w, we get that $\sum_{i=1}^{n} \max_{v \in [\varphi_i]} s(v, w) < ns(v_1, v_1)$. In conclusion, $s(\vec{\varphi}, w) < s(\vec{\varphi}, v_1) = s(\vec{\varphi}, v_2)$.

Since the satisfaction measures s_{AV}, s_{PAV} or s_{bPAV} satisfy none of the properties S_{1-3}, Theorem 5.9 implies that the approval-based operators Δ^{AV}, Δ^{PAV} or Δ^{bPAV} do not satisfy postulate M_2. On the other hand, the satisfaction measures s_{hD} and s_{hH} do satisfy properties S_{1-3}, showing that the corresponding operators satisfy postulate M_2.

As mentioned, we do not require satisfaction measures to be symmetric and, indeed, s_{bPAV} is not symmetric (though the other satisfaction measures are). The following result shows that, in the presence of postulate M_2, symmetry is connected to postulate M_4.

Theorem 5.10

If a satisfaction-based merging operator Δ^s satisfies postulate M_2, then Δ^s satisfies postulate M_4 if and only if s also satisfies property S_4 (i.e., is symmetric).

Proof

Take a merging operator Δ^s that satisfies postulate M_2. By Theorem 5.9, this implies that the satisfaction measure s satisfies properties S_{1-3}.

("$\Rightarrow$") Suppose that Δ^s satisfies postulate M_4 but that s is not symmetric, i.e,. there exist interpretations v_1 and v_2 such that $s(v_1, v_2) \neq s(v_2, v_1)$, Take, then, a profile $\vec{\varphi} = (\varphi_1, \varphi_2)$, with $[\varphi_1] = \{v_1\}$ and $[\varphi_2] = \{v_2\}$, and a constraint μ such that $[\mu] = \{v_1, v_2\}$. We get that $s(\vec{\varphi}, v_1) = s(v_1, v_1) + s(v_2, v_1)$ and $s(\vec{\varphi}, v_2) = s(v_1, v_2) + s(v_2, v_2)$. From property S_3 we have that $s(v_1, v_1) = s(v_2, v_2)$, and by postulate M_4 we get that $s(\vec{\varphi}, v_1) = s(\vec{\varphi}, v_2)$. Thus, $s(v_1, v_2) = s(v_2, v_1)$, which is a contradiction.

("$\Leftarrow$") We assume that s is symmetric and set out to show that Δ^s satisfies postulate M_4. First of all, notice that if s satisfies property S_4, then properties S_1 and S_2 coincide. Second, we have that s satisfies property S_3, and thus the satisfaction of

an interpretation with itself is the same across the entire universe. Let us denote $s(v, v) = k$, for $v \in \mathcal{U}$.

Suppose now that Δ^s does not satisfy postulate M_4. This implies that there exist two formulas φ and φ', and an interpretation $v^* \in [\varphi]$ such that $s((\varphi, \varphi'), v^*) > s((\varphi, \varphi'), v_j)$, for all $v_j \in [\varphi']$, which is further unpacked as saying that:

$$s(\varphi, v^*) + s(\varphi', v^*) > s(\varphi, v_j) + s(\varphi', v_j), \tag{5.6}$$

for all $v_j \in [\varphi']$.

Next, we have that $s(\varphi, v^*) = \max_{v_i \in [\varphi]} s(v_i, v^*)$. But, since $v^* \in [\varphi]$ and s satisfies property S_2, we get that $s(\varphi, v^*) = s(v^*, v^*) = k$. Analogously, we have that $s(\varphi', v_j) = s(v_j, v_j)$, for all $v_j \in [\varphi_2]$. Plugging this into Inequality 5.6 and simplifying, we have that:

$$s(\varphi', v^*) > s(\varphi, v_j),$$

for all $v_j \in [\varphi']$. This means that:

$$\max_{v_i \in [\varphi']} s(v_i, v^*) > \max_{v_i \in [\varphi]} s(v_i, v_j),$$

for all $v_j \in [\varphi']$. Suppose $\max_{v_i \in [\varphi']} s(v_i, v^*) = s(v^{**}, v^*)$, for some $v^{**} \in [\varphi']$. Then we get that:

$$s(v^{**}, v^*) > \max_{v_i \in [\varphi]} s(v_i, v_j),$$

for all $v_j \in [\varphi']$, which implies that:

$$s(v^{**}, v^*) > s(v^*, v^{**}),$$

which is a contradiction, since we have assumed that s is symmetric.

Since the satisfaction measures s_{hD} and s_{hH} are symmetric and, as implied by Theorem 5.9, satisfy properties S_{1-3}, we obtain by Theorem 5.10 that they also satisfy postulate M_4. Together with Proposition 5.17, this yields the full picture for the binary satisfaction-based operators Δ^{hH} and Δ^{hD}.

Proposition 5.18

The operators Δ^{hH} and Δ^{hD} satisfy postulates M_{0-8}.

Proof

For the operator Δ^{hD}, Theorem 5.8 gives us that it is equivalent to the distance-based operator Δ^{D}, known to satisfy postulates M_{0-8}.

For the operator Δ^{hH}, it already follows from Proposition 5.17 it satisfies postulates M_{0-1}, M_3 and M_{5-8}. For postulates M_2 and M_4, notice that the satisfaction measure s_{hH} satisfies properties S_{1-4}. This implies, by Theorems 5.9 and 5.10, that Δ^{hH} satisfies postulates M_2 and M_4.

For the approval-based operators, satisfaction of postulates M_2 and M_4 is clarified by another perspective on satisfaction measures. A satisfaction measure s is a *counting index* if there exists a function $\sigma\colon \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, called *the witness of* s, such that $\sigma(0,0) = 0$ and $s(v,w) = \sigma(|v \cap w|, |w|)$, for any interpretations v and w. Theorem 5.11 shows that counting indices do not fit with postulate M_2.

Theorem 5.11 ([Haret et al., 2020])

If s is a counting index, the satisfaction-based merging operator Δ^s does not satisfy postulate M_2.

It is straightforward to see that the approval-based satisfaction measures introduced in this section are counting indices. Thus, by Theorem 5.11, none of the operators they generate satisfies postulate M_2. For postulate M_4, however, the situation is different. Example 5.15 shows that the Δ^{AV} and Δ^{PAV} operators do not satisfy postulate M_4, though Δ^{bPAV} manages to evade the counter-example. In fact, it turns out that not only does the operator Δ^{bPAV} satisfy postulate M_4, but a much stronger result can be shown: it is the *only* operator based on a counting index that does so.

Theorem 5.12 ([Haret et al., 2020])

If Δ^s is a satisfaction-based merging operator such that s is a counting index with σ as witness, extends PAV and satisfies postulate M_4, then $\sigma(x,y) = 2h(x) - h(y)$, for any $x, y \in \mathbb{R}$.

It deserves emphasis that Δ^{bPAV} manages to satisfy postulate M_4 even though s_{bPAV} is not a symmetric satisfaction measure: since Δ^{bPAV} does not satisfy M_2, Theorem 5.10 does not apply. Indeed, none of the approval-based operators manages to satisfy both M_2 and M_4. This suggests that there is a trade-off between the kind of proportionality these operators stand for and the satisfaction of M_2 and M_4.

It is relevant that approval-based operators can consider interpretations of various sizes: reflection on Examples 5.15 and 5.17 shows that they are at the root of the problematic situations. Interestingly, it turns out that fixing the size of the models of the constraint μ yields merging operators that behave well with respect to the M postulates.

Theorem 5.13

If all models of the constraint μ have some fixed size k, then the approval-based merging operators Δ^{AV}, Δ^{PAV} and Δ^{bPAV} satisfy all postulates M_{0-8}.

Proof

It already follows from Proposition 5.17 that the operators Δ^s, for $s \in \{\mathrm{AV}, \mathrm{PAV}, \mathrm{bPAV}\}$ satisfy postulates M_{0-1}, M_3 and M_{5-8}. When the models of μ have fixed size k. All that is left to show is that these operators also satisfy postulates M_2 and M_4. The simplest way to see this is to notice that if we restrict the satisfaction function to take only interpretations of fixed size k in the second position, then the satisfaction functions s, for $s \in \{\mathrm{AV}, \mathrm{PAV}, \mathrm{bPAV}\}$, satisfy properties S_{1-4}. Therefore, by Theorems 5.9 and 5.10, they also satisfy postulates M_2 and M_4.

Two types of proportionality

Here we formalize two notions of proportionality, arising out of two different ways of conceptualizing satisfaction with respect to a possible outcome. For the sake of clarity, we define these notions for rather restricted profiles.

A formula φ is *complete* if it has exactly one model, and a profile $\vec{\varphi}$ is complete if all its formulas are complete. We write $\vec{\varphi} = (v_1, \ldots, v_n)$ to denote the complete profile with $[\varphi_i] = \{v_i\}$, for all $i \in \{1, \ldots, n\}$. A complete profile $\vec{\varphi} = (v_1, \ldots, v_n)$ is *simple* if $v_1 \cup \cdots \cup v_n = A$, and either $v_i = v_j$ or $v_i \cap v_j = \emptyset$, for every $i, j \in \{1, \ldots, n\}$.In the context of ABC voting, such profiles are referred to as party-list profiles [Lackner and Skowron, 2018b]: the term *party* refers here to political parties in parliamentary elections, where voters have to approve all candidates of one party and cannot vote for sub- or supersets.

A complete profile $\vec{\varphi} = (v_1, \ldots, v_n)$ is ℓ-*simple* if it is simple and $|\{v_1, \ldots, v_n\}| = \ell$, i.e., $\vec{\varphi}$ contains ℓ distinct sets. If $v_1, \ldots, v_\ell$ constitutes a partition of A, and $p_1, \ldots, p_\ell$ are positive integers, we write $(v_1^{p_1}, \ldots, v_\ell^{p_\ell})$ to denote the ℓ-simple profile:

$$(\underbrace{v_1, \ldots, v_1}_{p_1 \text{ times}}, \underbrace{v_2, \ldots, v_2}_{p_2 \text{ times}}, \ldots, \underbrace{v_\ell, \ldots, v_\ell}_{p_\ell \text{ times}}).$$

If $\vec{\varphi} = (v_1^{p_1}, \ldots, v_\ell^{p_\ell})$ is an ℓ-simple profile with $\sum_{i=1}^{\ell} p_i = n$, we say that $\vec{\varphi}$ is k-*integral* if $\frac{k \cdot p_i}{n}$ is an integer, for every $i \in \{1, \ldots, \ell\}$. Intuitively, for a model w of μ of size k, the fraction $\frac{k \cdot p_i}{n}$ denotes the intended satisfaction if proportionality is taken into account: out of the k atoms selected, the share of group i should be the relative size of the group.

We propose two proportionality postulates, intended to apply to any simple profile $\vec{\varphi} = (v_1^{p_1}, \ldots, v_\ell^{p_\ell})$. and constraint μ_k whose models are all interpretations of size k:

	s_{PAV}			s_{hH}		
	$3 \cdot a_1a_2a_3a_4a_5a_6$	b_1b_2	sum	$3 \cdot a_1a_2a_3a_4a_5a_6$	b_1b_2	sum
$a_1a_2a_3a_4$	$3 \cdot h(4)$	$h(0)$	6.25	$3 \cdot h(6)$	$h(2)$	**8.85**
$a_1a_2a_3b_1$	$3 \cdot h(3)$	$h(1)$	**6.5**	$3 \cdot h(4)$	$h(4)$	8.33
$a_1a_2b_1b_2$	$3 \cdot h(2)$	$h(2)$	6.0	$3 \cdot h(2)$	$h(2)$	6.95
...						

Table 5.16: Satisfaction s_{PAV} and s_{hH}, as well as the aggregates satisfactions, for profile $\vec{\varphi}$ and constraint μ.

(M$_{\mathsf{CPROP}}$) For any $k \in \{1, \dots, m\}$ and $w \in [\Delta\mu_k(\vec{\varphi})]$, it holds that if $\vec{\varphi}$ is k-integral and $|v_j| \geq \frac{k \cdot p_j}{n}$ for each j, $1 \leq j \leq l$, then $|v_i \cap w| = \frac{k \cdot p_i}{n}$, for all $i \in \{1, \dots, \ell\}$.

(M$_{\mathsf{BPROP}}$) If $\vec{\varphi} = (v_1^{p_1}, v_2^{p_2})$ is simple and there is a $w \in [\mu]$ such that $m - d_{\text{H}}(v_i, w) = \frac{m \cdot p_i}{n}$ for $i \in \{1, 2\}$, then this equality holds for all $w' \in [\Delta\mu(\vec{\varphi})]$.

We refer to M$_{\mathsf{CPROP}}$ and M$_{\mathsf{BPROP}}$ as postulates of *weak classical proportionality* and *weak binary proportionality*, respectively, as they refer to different sources of satisfaction. Postulate M$_{\mathsf{CPROP}}$ talks about *classical satisfaction*, in which agent i's satisfaction with an interpretation w is given by $|v_i \cap w|$, just like the satisfaction with a committee in an ABC election is measured by the number of approved committee members. This is the kind of satisfaction notion typically used in a social choice context. Postulate M$_{\mathsf{BPROP}}$ talks about *binary satisfaction*, in which agent i's satisfaction with w is given by $m - d_{\text{H}}(v_i, w)$, i.e., by the closeness between v_i and w. This type of satisfaction follows from a logical viewpoint where positive and negative variable assignments are treated equally. This approach is better suited to deal with interpretations of varying sizes than the classical one, and thus postulate M$_{\mathsf{BPROP}}$ allows such interpretations to be selected.

Intuitively, both postulates stipulate 'shares' groups of agents shall receive, under a classical or binary viewpoint, that meet proportionality based on the relative size of the groups, in case the profile satisfies the specified conditions. For postulate M$_{\mathsf{CPROP}}$ we restrict the constraint to μ_k, with k atoms to be distributed proportionally by each solution w (like for ABC elections). Postulate M$_{\mathsf{BPROP}}$ states that in the presence of at least one admissible $w \in [\mu]$ that meets the proportionality requirements, *all* solutions shall meet said requirements (otherwise μ permits no proportional solution). Note that if $\vec{\varphi} = (v_1^{p_1}, v_2^{p_2})$ satisfies the conditions of M$_{\mathsf{BPROP}}$, then $\vec{\varphi}$ is m-integral, and the binary satisfaction of v_1 and v_2 adds up to m, i.e., $m - d_{\text{H}}(v_1, w) + m - d_{\text{H}}(v_2, w) = m$. Postulate M$_{\mathsf{BPROP}}$ demands that this total satisfaction m is split proportionally.

Example 5.18: Classical and binary proportionality

For the set of atoms $A \cup B$, with $A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ and $B = \{b_1, b_2\}$, take the simple profile $\vec{\varphi} = (v_1^3, v_2^1)$, with $v_1 = a_1a_2a_3a_4a_5a_6$ and $v_4 = b_1b_2$, and a constraint

μ_4, with models of size 4. Since the number of formulas in $\vec{\varphi}$ is 4 (i.e., $n = 4$), it is easy to see that the profile $\vec{\varphi}$ is 4-integral as well as 8-integral (the latter is needed because $m = 8$ in this case).

We would like to understand what kind of interpretations would be chosen by a merging operator Δ that satisfies postulate $\mathrm{M}_{\mathsf{CPROP}}$ and $\mathrm{M}_{\mathsf{BPROP}}$, respectively. To do this, we can use the equalities presented in these postulates to infer properties of an optimal outcome, i.e., an interpretation w such that $w \in [\Delta_\mu(\vec{\varphi})]$. In these equalities, we are basically treating w as an unknown and solving for it.

Assume, first, we are working with a merging operator Δ that satisfies postulate $\mathrm{M}_{\mathsf{CPROP}}$. Thus, for $k = 4$, postulate $\mathrm{M}_{\mathsf{CPROP}}$ tells us that if $w \in [\Delta_{\mu_4}(\vec{\varphi})]$, then it holds that:

$$\begin{aligned} |v_1 \cap w| &= \frac{4 \cdot 3}{4} \\ &= 3, \end{aligned}$$

and:

$$\begin{aligned} |v_2 \cap w| &= \frac{4 \cdot 1}{4} \\ &= 1. \end{aligned}$$

Thus, from the standpoint of classical proportionality, an optimal outcome of size 4 reflects the proportion of agents that approve atoms within it, and it would contain three variables from A and one from B, e.g., the interpretation $w = a_1a_2a_3b_1$.

Assume, however, that we are working with a merging operator Δ that satisfies postulate $\mathrm{M}_{\mathsf{BPROP}}$. Postulate $\mathrm{M}_{\mathsf{BPROP}}$ tells us that if there exists an interpretation $w' \in [\Delta_{\mu_4}(\vec{\varphi})]$ that satisfies the equality:

$$m - d_{\mathrm{H}}(v_i, w) = \frac{m \cdot p_i}{n},$$

for $i \in \{1, 2\}$, then every interpretation in $[\Delta_\mu(\vec{\varphi})]$ satisfies this equality. Suppose there exists such an interpretation w. The equality requires that:

$$\begin{aligned} 8 - d_{\mathrm{H}}(v_1, w) &= \frac{8 \cdot 3}{4} \\ &= 6, \end{aligned}$$

and:

$$\begin{aligned} 8 - d_{\mathrm{H}}(v_2, w) &= \frac{8 \cdot 1}{4} \\ &= 2. \end{aligned}$$

This implies that $d_{\mathrm{H}}(v_1, w) = 2$ and $d_{\mathrm{H}}(v_2, w) = 6$. Among all possible interpretations, the ones that satisfy these conditions are:

(a) interpretations of size 4, consisting of four atoms from A, e.g., $a_1a_2a_3a_4$;

(b) interpretations of size 6, consisting of five atoms from A and one from B, e.g., $a_1a_2a_3a_4a_5b_1$;

(c) the interpretation consisting of all atoms from A and all atoms from B, i.e., $a_1a_2a_3a_4a_5a_6b_1b_2$.

Thus, postulate $\mathrm{M}_{\mathsf{BPROP}}$ says that if at least one of these interpretations are in mods μ, then the interpretations that make it into the result are all from the same list.

Note that from the standpoint of binary proportionality, it makes sense to select among interpretations of varying sizes, as the satisfaction notion is calibrated to take into account the differences that arise. Note also that if the constraint is restricted to interpretations of size 4 (i.e., the constraint is μ_4), then only interpretations of type (a) get selected. In this setup, an interpretation such as $w = a_1a_2a_3b_1$ provides less satisfaction to $\vec{\varphi}$ than interpretations containing only atoms from A, such as $w' = a_1a_2a_3a_4$. This is because for agents v_1, v_2 and v_3 the exclusion of the desired atom a_4 at the expense of the undesired atom b_1 (when going from w' to w) incurs double the penalty as in the case of classical proportionality.

The quantity $|v_i \cap w|$ in $\mathrm{M}_{\mathsf{CPROP}}$ is indicative of notions of classical satisfaction, while the quantity $m - d_{\mathrm{H}}(v_i, w)$ in $\mathrm{M}_{\mathsf{BPROP}}$ is indicative of binary satisfaction. The operators Δ^{PAV} and Δ^{bPAV} are representatives of the former notion and the operator Δ^{hH} is representative of the latter. Notice that under the constraint μ_4 the operators Δ^{PAV} and Δ^{bPAV} select interpretations that have three atoms from A and one from B, e.g., $a_1a_2a_3b_1$, while Δ^{hH} selects interpretations that have only atoms from A, e.g., $a_1a_2a_3a_4$. See Table 5.16 for an illustration.

Example 5.18 shows that classical and binary proportionality may require different interpretations to be selected on the same input. Thus, even though our notions of proportionality apply only to simple profiles, they set up a clear boundary for distinguishing among the different merging operators.

Theorem 5.14 ([Haret et al., 2020])

The merging operators Δ^{PAV} and Δ^{bPAV} satisfy postulate $\mathrm{M}_{\mathsf{CPROP}}$, Δ^{hH} satisfies postulate $\mathrm{M}_{\mathsf{BPROP}}$, while $\Delta^{\mathrm{H,sum}}$, $\Delta^{\mathrm{H,leximax}}$, Δ^{hD} and Δ^{AV} satisfy neither $\mathrm{M}_{\mathsf{CPROP}}$ nor $\mathrm{M}_{\mathsf{BPROP}}$.

The proposed merging operators Δ^{PAV} and Δ^{bPAV} are representative of the notion of classical proportionality, while Δ^{hH} is representative for binary proportionality. Theorem 5.15 shows that these notions are thoroughly incompatible.

Theorem 5.15 ([Haret et al., 2020])

There is no merging operator that satisfies M_1 and both $\mathrm{M}_{\mathsf{CPROP}}$ and $\mathrm{M}_{\mathsf{BPROP}}$.

5.6 Related work

Work trying to connect belief merging and social choice can be divided along two lines. One direction of research views voting as a merging task [Eckert and Pigozzi, 2005, Gabbay et al., 2007], an approach which fits into the larger program of finding suitable logics in which to represent preferences and embed aggregation problems stemming from (computational) social choice [Chevaleyre et al., 2008, Endriss, 2011].

Here, however, we take a different approach and look at merging from a voting perspective, with the aim of using the rich set of criteria developed to analyze voting rules in order to classify existing merging operators. Work in this area has focused on impossibility results in the style of Arrow's theorem [Maynard-Zhang and Lehmann, 2003, Konieczny and Pino Pérez, 2005, Chopra et al., 2006, Díaz and Pino Pérez, 2017], strategyproofness [Everaere et al., 2007, Díaz and Pino Pérez, 2018] and the analysis of additional desirable properties for merging operators, e.g., egalitarian properties [Everaere et al., 2014]. Notwithstanding, we find that the social choice literature on voting features many other leads that are relevant to the aggregation of information in the context of merging. Thus, in Section 3.4 it was already mentioned that existing merging operators $\Delta^{\mathrm{H,\,sum}}$ and $\Delta^{\mathrm{H,\,leximax}}$ embody two different strategies, with the former being more majoritarian while the latter being more egalitarian: in this chapter we follow this line of reasoning further and investigate ways of looking at merging operators that improve upon the basic distinction between majoritarian and egalitarian operators.

The work on manipulation of belief merging operators that is closest to ours is [Everaere et al., 2007]. The setup there differs from the one we work with, in that satisfaction indices in [Everaere et al., 2007] are not based on skeptical or credulous acceptance but on the models that the strategic agent and the result have in common. To highlight this difference, note that under the indices in [Everaere et al., 2007] the strategic agent in Example 5.6, i.e., Academy member 2, would be equally (dis)satisfied with both the truthful result $\Delta_\mu(\vec{\varphi}^{\mathsf{t}})$ and the manipulated result $\Delta_\mu(\vec{\varphi}^{\mathsf{f}})$, since ${\varphi_2}^{\mathsf{t}}$ shares exactly one model with both. However, under our interpretation of the indices, $\Delta_\mu(\vec{\varphi}^{\mathsf{f}})$ ends up delivering a better result for the strategic agent than $\Delta_\mu(\vec{\varphi}^{\mathsf{t}})$, as under $\Delta_\mu(\vec{\varphi}^{\mathsf{f}})$ the atoms a and b are guaranteed to be in the result, and there is a sense in which this is satisfactory for the strategic agent, as a and b are atoms that they skeptically accept. In a further difference with [Everaere et al., 2007] and [Díaz and Pino Pérez, 2018] we also show results for manipulation with respect to an atom, which is not based on indices.

Belief merging invites comparison to multiwinner elections [Faliszewski et al., 2017a], combinatorial voting [Lang and Xia, 2016] and Judgment Aggregation [Endriss, 2016, Baumeister et al., 2017]. We mention here that our use of acceptance notions and satisfaction

indices, the compact encoding of sets of interpretations (agents' "top candidates") as propositional formulas, and the fact that we do not require the output to be of a specific size suggest that existing results in this area are not directly applicable to our setting. Our work does intersect with social choice in the special case when the profile is complete and the number of bases is odd. In this case the aggregation problem corresponds to a Judgment Aggregation problem, with the operator $\Delta^{\mathrm{H,\,sum}}$ and the constraint μ set to $\top$ delivering the majority opinion on the atoms (considered as issues): this corresponds to the observation made in the Social Choice literature [Brams et al., 2007] that the majority opinion minimizes the sum of the Hamming distances to voters' approval ballots. Our strategy-proofness result for $\Delta^{\mathrm{H,\,sum}}$ with the constraint μ set to $\top$ dovetails neatly with a similar result in Judgment Aggregation [Baumeister et al., 2017, Endriss, 2016], though our treatment is slightly more general, as it accommodates both an even and an odd number of bases.

The literature on belief merging suggests other properties concerned, in some way or another, with fairness of merging operators [Konieczny and Pino Pérez, 2002, Everaere et al., 2010a, Everaere et al., 2014], and there is the question of how the operators we introduced in Section 5.5 stand up against these postulates. Regarding the majority postulate $\mathrm{M}_{\mathsf{MAJ}}$ and the arbitration postulate $\mathrm{M}_{\mathsf{ARB}}$ in [Konieczny and Pino Pérez, 2011], and also presented in Section 3.4, we mention, first of all, that the proposed operators do satisfy postulate $\mathrm{M}_{\mathsf{MAJ}}$: a large enough majority will eventually tilt the result in its favor. However, our rules are not majoritarian in the sense that a 51% majority can dictate the outcome. Second, the proposed operators do not satisfy postulate $\mathrm{M}_{\mathsf{ARB}}$: a counterexample to the corresponding semantic property is obtained with the set of atoms $A = \{a, b, c, d, x, y, z, t, u, v\}$ and profile $\vec{\varphi} = (\varphi_1, \varphi_2)$, with $[\varphi_1] = \{abcd\}$, $[\varphi_2] = \{xyzt\}$ and constraint $[\mu] = \{axuv, xyzt, abcd\}$.

The work in [Everaere et al., 2014] is largely orthogonal to our work on proportionality: the median and cumulative sum operators there do not deliver proportional results, and our operators do not satisfy properties (SHE) and (PD). Finally with respect to the operators in [Everaere et al., 2010a], the guiding postulate (Disj) there is incompatible with proportionality as we formalize it. For instance, in Example 5.14, postulate (Disj) requires the result to be a subset of $\{a_1a_2a_3a_4a_5, b_1b_2b_3b_4b_5\}$, which runs against all the proportionality requirements we expect.

Finally, we mention that some of the work done here can be ported back to the social choice literature. By setting μ to $\top$, merging operators can be seen as ABC voting rules with variable committee sizes [Kilgour, 2016, Faliszewski et al., 2017b]. It is easy to see that the AV and PAV operators are not sensible in this context, as $w = A$ (i.e., setting all atoms to true) is always an optimal model. However, the bounded PAV operator Δ^{bPAV} and the harmonic Hamming operator Δ^{hH} present themselves as novel additions to this framework, being *proportional* ABC rules with variable committee sizes.

5.7 Conclusion

Seeing merging as a collective decision mechanism has allowed us to ask what kind of fairness properties are suitable for merging operators, beyond the ones enshrined in postulates M_{0-8}. In this chapter we have investigated some desirable properties for belief merging operators, most of them suggested by the social choice literature: insensitivity to syntax in Section 5.1, collective efficiency in Section 5.2, responsiveness in Section 5.3, vulnerability to manipulation in Section 5.4 and proportionality in Section 5.5.

In Sections 5.1, 5.2 and 5.3 we put forward an assortment of properties originally thought to apply to voting rules [Zwicker, 2016, Baumeister and Rothe, 2016], and found that some of the properties proposed follow postulates M_{0-8}, some are at odds with these postulates, whereas others are only satisfied by certain existing operators.

In Section 5.4 we proposed a two-sided approach to manipulation of skeptical or credulous consequences: (*i*) by considering what we call *constructive* and *destructive* manipulation, where the aim is to usher a desired atom into (or out of) the skeptical or credulous consequences, and (*ii*) by adapting an earlier approach to manipulation [Everaere et al., 2007] that utilizes satisfaction indices to quantify the (dis)satisfaction of agents with respect to the merged outcomes; our contribution here consists in proposing new indices. We showed that all the main merging operators are manipulable, even when enforcing restrictions that yielded non-manipulability in earlier works [Everaere et al., 2007]. The sole exception is the case when merging is done using only complete formulas (i.e., having exactly one model) the integrity constraint is set to $\top$ and the merging operator is $\Delta^{\mathrm{H,\,sum}}$, under our new dissatisfaction indices. On the question of how an agent can manipulate, we looked at general approaches to influencing the aggregation procedure by promoting or demoting interpretations. Further, we showed that manipulation under skeptical consequences can be carried out by the strategic agent submitting a complete base, suggesting that manipulation does not require sophisticated propositional structures to succeed.

For defining proportional belief merging operators in Section 5.5, we relied on the Proportional Approval Voting (PAV) rule, studied in multiwinner voting scenarios and known to satisfy particularly strong proportionality requirements [Aziz et al., 2017]. Based on the PAV rule, we introduced a series belief merging operators, the most important of which turned out to be the PAV operator Δ^{PAV}, the bounded PAV operator Δ^{bPAV} and the harmonic Hamming operator Δ^{hH}. All these operators fall into the class of *satisfaction-based operators*, introduced by us as an alternative to the standard way of representing merging operators, which is distance-based. We looked at the proposed belief merging operators from three perspectives. Firstly, the operators were placed against the standard belief merging postulates M_{0-8}. We showed that any belief merging operator directly extending PAV cannot be compatible with all of the postulates: in particular, such an operator will not satisfy postulate M_2. We also provided a characterization of operators that fail postulate M_2, based on properties of the ranking a satisfaction-based operator induces, which provides an alternative view on why the PAV approach to

proportionality is inconsistent together with postulate M_2. At the same time, we saw that the bounded PAV operator can be characterized as the only merging operator (of a certain natural class) that extends PAV and satisfies all other postulates. While the harmonic Hamming operator is defined via the harmonic sum used by PAV, it does not generalize PAV. Thus, the aforementioned impossibility does not hold; indeed, the harmonic Hamming operator satisfies all standard postulates M_{0-8}.

Secondly, we introduced two basic proportionality postulates. Postulate M_{CPROP}, concerning *classical proportionality*, is the kind of proportionality requirement typically studied in social choice settings, in particular in the apportionment setting [Balinski and Young, 1982]. This notion is based on the assumption that agents derive utility from positive occurrences, i.e., from approved candidates being selected in the collective choice. Postulate M_{BPROP}, concerning *binary proportionality*, is closer to the logical nature of belief merging. Here, no difference is made between positive and negative agreement: the agents' utility derives from the (Hamming) distance between their preferences and the collective choice. It was observed that these two notions are mutually exclusive and contradict each other. Furthermore, we showed by example that established belief merging operators satisfy neither of these two postulates. In contrast, the aforementioned PAV and bounded PAV operators satisfy classical proportionality and the harmonic Hamming operator satisfies binary proportionality.

On a general note, it emerged that the PAV operator is biased towards larger interpretations, i.e., it tries to make everyone happy by setting as many atoms true as possible. If the result is assumed to be made up of interpretations of fixed size then this is not a problem; but if not, then the PAV operator is too greedy to be usable in practice, and either the bPAV or the hH operators are preferable. The bPAV operator is best used in contexts close to those studied in social choice, where positive literals are the main focus, whereas hH is likely to be more useful in logic-based approaches, where positive and negative literals count equally.

CHAPTER 6

Belief Change for Horn Formulas

In this chapter we look at revision and update for *Horn formulas*, a type of propositional formulas used to represent facts and rules, i.e., information of the type *if . . . , then. . . .* Horn formulas form a subset of the language $\mathcal{L}$ of propositional logic, and are therefore referred to as a *fragment of propositional logic*. Interest in the Horn fragment arises because of a number of salient features: important reasoning tasks, such as checking consistency, become tractable for Horn formulas, and the restrictions imposed on the language mirror widely used formalisms used in Knowledge Representation (KR), e.g., logic programming, databases and description logics. Thus, there are computational benefits of assuming that an agent's epistemic state is expressed by a Horn formula. The cost lies in the decreased expressivity, since not all propositional formulas can be recast as a Horn formula. Nonetheless, if the type of information we are working with lends itself to the format laid down by the Horn fragment, this is a tradeoff that, in many cases, is worth making,

Concern about practical aspects has led to an increase in efforts to understand belief change in formalisms that can lay claim to being useful in applications, e.g., fragments weaker than propositional logic and the Horn fragment in particular. These efforts have resulted in a model of considerable range and generality [Delgrande and Peppas, 2015, Delgrande et al., 2018], which we take here as reference point. Apart from its relevance to various KR formalisms, the role of the Horn fragment is as a belief change guinea pig, i.e., a testing ground for new approaches to belief change, before these get deployed in real world applications.

Research on belief change for Horn formulas is typically done with an eye towards finding appropriate postulates and deriving representation results in the same spirit as the representation results seen for propositional logic. In this chapter, where we look at revision and update for Horn formulas, the models we want to emulate are Theorems 3.5 and 3.6 for revision, and Theorems 3.7 and 3.8 for update. In keeping with the choice perspective developed in Chapter 3, we want to show that revision and update operators

applied to Horn formulas can be characterized as choice functions over preorders induced by the prior beliefs. The limited expressivity of the Horn fragment, however, means that familiar postulates and typical results break down if additional restrictions are not added.

Belief change in the Horn fragment requires that the agent's epistemic state i.e., the snapshot of the information it has in its 'head' at any given moment, is expressible as a Horn formula. In concrete terms this translates as saying that, at the very least, the prior and the posterior information, are Horn formulas. This leaves open the situation with respect to the new information, which can be Horn or not, depending on the source of information. In this chapter we look at two cases: in the first case, the new information is a propositional formula; in the second case, it is assumed to be a Horn formula. More concretely, we will first study what we call HPH-*revision*, in which the prior information φ and the posterior information $\varphi \circ \mu$ are both Horn formulas, but μ is allowed to be any propositional formula. Then we will move our attention to HHH-*revision* and HHH-*update*, in which the prior information, posterior information, as well as the new information are Horn formulas.

In the case of HPH-revision we will see that the mild assumption that μ is a propositional formula clashes systematically with the commonly accepted postulate R_2, as well as with the neutrality postulate $\mathrm{R}_{\mathsf{NEUT}}$ described in Section 4.1. We will show that postulate R_2 puts certain demands on the underlying language (e.g., that the conjunction of φ and μ is always expressible in it), which are not met in all scenarios that interest us. Thus, there is an unexpected payoff in looking, as we have done in Chapter 4, at weaker versions of the standard postulates and their semantic characterizations.

The case of HHH-revision with postulates R_{1-6} largely corresponds to work that has already been done [Delgrande and Peppas, 2015, Delgrande et al., 2018], and here we present it only in the interest of drawing a coherent picture, and to lay down the groundwork for our contribution, which shows that the existing results can be extended to the weaker revision postulates R_{1-5} and R_{7-8}.

For HHH-update we show that, as with existing work on Horn revision, standard results do not generalize in a straightforward way. First, special care must be taken when stating postulates, as the limited expressibility of the Horn fragment makes formulation of familiar intuitions either cumbersome or impossible: since the Horn fragment is not closed under disjunction, certain postulates must be weakened, but this then results in the possibility that Horn operators are represented by undesirable types of preorders on outcomes. This difficulty is reminiscent of problems encountered when characterizing Horn revision using total preorders [Delgrande and Peppas, 2015]. However, since our aim is to capture Horn update operators characterizable with partial (as well as total) preorders, these problems are compounded and require new ideas. We handle this issue by adding new postulates whose effect is felt in the Horn fragment, but which follow from the standard postulates in propositional logic. Second, it turns out that standard operators proposed in the literature (e.g., Forbus' and Winslett's operators) do not meet it and a special restriction, called here *Horn compliance*, must be placed on any acceptable operator.

6.1 The Horn fragment

At its most general, a fragment $\mathcal{L}_\star$ of propositional logic is a set $\mathcal{L}_\star \subseteq \mathcal{L}$ of propositional formulas. In this chapter we are mainly interested in the Horn fragment.

Recall that if A is the set of propositional atoms, then a *literal* l is either an atom in A or its negation. If l is an atom, then l is a *positive literal* and if l is a negated atom then it is a *negative literal*. A *propositional clause* is a disjunction of literals, and a *Horn clause* is a clause that contains at most one positive literal. A *Horn formula* φ is a propositional formula that is a conjunction of Horn clauses. The *Horn fragment* $\mathcal{L}_{\text{Horn}}$ is the set of all Horn formulas. The semantics of Horn formulas is the same as for propositional formulas.

Example 6.1: Horn formulas

If the set of atoms is $A = \{a, b, c\}$, then $\varphi_1 = \neg a$, $\varphi_2 = \neg a \vee c$, $\varphi_3 = \neg a \vee \neg b \vee c$, $\varphi_4 = \neg a \wedge (\neg a \vee c)$ are all Horn formulas. The formula $\varphi_5 = a \vee b$, however, is not.

Note that φ_1, φ_2, φ_3 and φ_4 are semantically equivalent to $\varphi'_1 = a \to \bot$, $\varphi'_2 = a \to c$, $\varphi'_3 = (a \wedge b) \to c$, $\varphi'_4 = (a \to \bot) \wedge (a \to c)$.

In Example 6.1, Horn formulas could be rewritten as statements of facts (i.e., single literals) or rules involving facts (i.e., conditional *if . . . , then . . .* statements). This is a useful way of thinking about Horn formulas, and is the feature that makes them useful to many KR formalisms.

We have so far characterized Horn formulas only syntactically, but Chapter 3 has prepared us to expect that belief change operators do not care much about syntax. Therefore, we want to understand Horn formulas at the semantic level as well: in particular, what it takes for a set of interpretations $\mathcal{W}$ to be the set of models of some Horn formula φ. Formally, the link between the syntax and the semantics of Horn formulas is provided by a *closure operator* Cl, which is a function $\text{Cl}\colon 2^{\mathcal{U}} \to 2^{\mathcal{U}}$, taking a set $\mathcal{W}$ of interpretations as input and returning a set $\text{Cl}(\mathcal{W})$ of interpretations as output. If $\mathcal{W}$ is a set of interpretations and $\text{Cl}_\star$ is a closure operator, then $\mathcal{W}$ *is* $\star$*-closed* if $\text{Cl}_\star(\mathcal{W}) = \mathcal{W}$. Intuitively, a closure operator Cl transforms a set of interpretations in a certain way, with $\star$-closed sets being left unchanged. We will use this transformation to characterize the semantics of a fragment.

Since we will be looking at only the Horn fragment in this chapter we will not make much of the properties expected to hold of a closure operator in general, except to say that if $\text{Cl}_\star$ is a closure operator, then $\mathcal{L}_{\text{Horn}}$ *is characterized by* $\text{Cl}_\star$ if, for any formula φ in $\mathcal{L}_{\text{Horn}}$ and any set of interpretations $\mathcal{W}$, it holds that:

(*a*) $\text{Cl}_\star([\varphi]) = [\varphi]$.

(*b*) If $\text{Cl}_\star(\mathcal{W}) = \mathcal{W}$, then there exists a formula φ in $\mathcal{L}_\star$ such that $[\varphi] = \mathcal{W}$.

Intuitively, the Horn fragment $\mathcal{L}_{\text{Horn}}$ is characterized by a closure operator $\text{Cl}_\star$ if the models of every Horn formula are closed under the operator $\text{Cl}_\star$ and any set of interpretations $\mathcal{W}$ that is closed under $\text{Cl}_\star$ is the set of models of some Horn formulas.

The question, now, is what closure operator characterizes the Horn fragment. We raise this question only to answer it: consider the *intersection function* ints, which is a function $\text{ints}\colon 2^{\mathcal{U}} \to 2^{\mathcal{U}}$, defined as:

$$\text{ints}(\mathcal{W}) = \{w_1 \cap w_2 \mid w_1, w_2 \in \mathcal{W}\}.$$

Intuitively, the intersection function ints adds to a set $\mathcal{W}$ of interpretations all the interpretations obtained by intersecting interpretations in $\mathcal{W}$. This is a function we will want to iterate. Thus, if $\mathcal{W}$ is a set of interpretations and $k \geq 0$ is an integer, then $\text{ints}^0(\mathcal{W}) = \mathcal{W}$ and $\text{ints}^{k+1}(\mathcal{W}) = \text{ints}^k(\mathcal{W})$. Clearly, iterating the ints function on a finite set $\mathcal{W}$ of interpretations reaches a fixed point, i.e., there exists an integer k such that $\text{ints}^{k+i}(\mathcal{W}) = \text{ints}^k(\mathcal{W})$, for any $i \geq 0$. We will denote by ints^* the fixed point of the intersection function ints, and define the *Horn closure operator* Cl_{Horn}, for any set of interpretations $\mathcal{W}$, as:

$$\text{Cl}_{\text{Horn}}(\mathcal{W}) = \text{ints}^*(\mathcal{W}).$$

A set of $\mathcal{W}$ interpretations is Horn-*closed* if $\text{Cl}_{\text{Horn}}\mathcal{W} = \mathcal{W}$. The answer to the question we started with is provided by the next result.

Proposition 6.1 ([McKinsey, 1943, Horn, 1951])

The Horn fragment $\mathcal{L}_{\text{Horn}}$ is characterized by the Cl_{Horn} closure operator.

Intuitively, Proposition 6.1 says that the semantic property characterizing Horn formulas is closure under intersection: a propositional formula φ is (or is equivalent to) a Horn formula if and only if the set $[\varphi]$ of its models is closed under intersection. Since the semantics of Horn formulas is more important to belief change operators than their syntax, we will subsequently be more loose in what we call a *Horn formula*: we will use the term to refer, more generally, to any formula whose set of models is closed under intersection, regardless of whether it belongs to $\mathcal{L}_{\text{Horn}}$ according to its proper definition. The rationale for this usage is that if the set of models of a propositional formula φ is closed under intersection, then φ is equivalent to some (properly) Horn formula φ^*, so we can always replace φ with φ^* if needed.

Example 6.2: Horn formulas and their semantics

For the set of atoms $A = \{a, b, c\}$ and interpretations ab and ac, we have that $\mathcal{W} = \{ab, ac\}$ is not Horn-closed, since the intersection of interpretations ab and ac (i.e., the interpretation a) is not in $\mathcal{W}$. Thus, there is no Horn formula $\varphi_{\mathcal{W}}$ that captures $\mathcal{W}$ exactly, in the sense that $[\varphi_{\mathcal{W}}] = \mathcal{W}$. However, the Horn-closure of $\mathcal{W}$, i.e., $\text{Cl}_{\text{Horn}}(\mathcal{W}) = \mathcal{W} \cup \{a\}$, does admit of such a formula, e.g., $\varphi_1 = a \wedge ((b \wedge c) \to \bot)$,

since, by definition, it is Horn-closed.

Consider, also, the Horn formula $\varphi_2 = b \wedge ((a \wedge c) \rightarrow \bot)$, with $[\varphi_2] = \{ab, bc, b\}$. We can see that $[\varphi_1 \wedge \varphi_2] = \{ab\}$, i.e., $\varphi_1 \wedge \varphi_2$ is also a Horn formula. On the other hand, we have that $[\varphi_1 \vee \varphi_2] = \{ab, bc, ac, a, b\}$ and thus $\varphi_1 \vee \varphi_2$ is not a Horn formula. It can happen, nonetheless, that the disjunction of two Horn formulas is another Horn formula. If $\varphi_3 = (a \rightarrow \bot) \wedge (c \rightarrow \bot)$, with $[\varphi_3] = \{\emptyset, b\}$, we have that $[\varphi_1 \vee \varphi_3] = \{ab, ac, a, b, \emptyset\}$, which is closed under intersection.

Note that, as Example 6.2 illustrates, not every set of interpretations corresponds directly to a Horn formula, i.e., unlike the language of propositional logic $\mathcal{L}$, the Horn fragment is not able to capture, or reach, all sets of interpretations. This limited expressiveness of the Horn fragment, while being a boon for computational matters, is what will make life difficult for belief change operators.

Example 6.2 also presents a case in which the conjunction of two Horn formulas is also a Horn formula: this holds more generally, i.e., the conjunction of any two (and, by extension, of any finite number of) Horn formulas is also a Horn formula. Rephrasing this fact as a mantra we can invoke whenever needed, we have that the Horn fragment $\mathcal{L}_{\text{Horn}}$ is *closed under conjunction*.

Before moving on, there is one notion that plays an important role in belief change and that still needs to be addressed: the proxy of a set of interpretations $\mathcal{W}$. In Chapters 3 and 4 we used the $\mathcal{L}$-proxy of $\mathcal{W}$ whenever we were in need of a propositional formula that applied exactly to the interpretations in $\mathcal{W}$; in choice terms, we were always able to present the choice function (i.e., revision operator) with a menu (i.e., formula) that consisted exactly of $\mathcal{W}$: if a comparison between interpretations w_1 and w_2 was needed, the revision operator could be queried using a propositional formula $\varepsilon_{1,2}$, with $[\varepsilon_{1,2}] = \{w_1, w_2\}$, and the result indicated the agent's assessment of which interpretation was more preferred. But Example 6.2 has just shown us that such a formula might not exist in the Horn fragment. What is there to be done?

The solution, standardly, is to find a Horn formula that approximately represents $\mathcal{W}$, i.e., even if it does not manage to capture $\mathcal{W}$, it comes sufficiently close to permit us to use it for the purposes of belief change. Thus, the $\mathcal{L}_{\text{Horn}}$*-proxy* $\varepsilon_{\mathcal{W}}$ *of* $\mathcal{W}$ is defined as a Horn formula such that $[\varepsilon_{\mathcal{W}}] = \text{Cl}_{\text{Horn}}(\mathcal{W})$. Note that $\varepsilon_{\mathcal{W}}$, thus defined, exists and generalizes the notion of an $\mathcal{L}$-proxy of $\mathcal{W}$. In particular, if $\mathcal{W} = \{w_1, w_2\}$ and is such that it is not Horn-closed, then an $\mathcal{L}_{\text{Horn}}$-proxy $\varepsilon_{\mathcal{W}}$ of $\mathcal{W}$ is such that $[\varepsilon_{\mathcal{W}}] = \{w_1, w_2, w_1 \cap w_2\}$.

Example 6.3: $\mathcal{L}_{\text{Horn}}$-proxies

For the set $\mathcal{W}$ of interpretations in Example 6.2, the Horn formula $\varphi_1 = a \wedge ((b \wedge c) \rightarrow \bot)$, with $[\varphi_1] = \{a, ab, ac\}$, serves as an $\mathcal{L}_{\text{Horn}}$-proxy.

6.2 Horn revision by propositional formulas

An HPH-*revision operator* $\circ$ is a function $\circ\colon \mathcal{L}_{\text{Horn}} \times \mathcal{L} \to \mathcal{L}_{\text{Horn}}$, taking as input a Horn formula φ and a propositional formula μ, and returning a Horn formula $\varphi \circ \mu$. We may assume that HPH-revision describes an agent who is able to process information that can take on any syntactic form, but is bound by its specifications to think only in terms of Horn formulas.

The natural next step now is to bring in postulates and assignments, and find a way to connect them. Ideally, we could use the standard postulates R_{1-6}, or any variants thereof: at the very least postulates R_1 and R_{3-4}, which we have singled out as the theoretical minimum that a revision operator should satisfy. And indeed, postulates R_1 and R_{3-5} can be adapted seamlessly to the Horn fragment. But consider what happens if we try to use postulate R_2.

Example 6.4: HPH-revision operators cannot satisfy postulate R_2

For the set of atoms $A = \{a, b\}$, consider the Horn formula $\varphi = \neg a \vee \neg b$ and the formula $\mu = a \leftrightarrow \neg b$. Note that $[\varphi] = \{\emptyset, a, b\}$ and $[\mu] = \{a, b\}$: since $\emptyset = a \cap b \notin [\mu]$, μ is not a Horn formula.

Clearly, $\varphi \wedge \mu$ is consistent and, moreover, $\varphi \wedge \mu \equiv \mu$. However, $[\varphi \wedge \mu] = \{a, b\}$, which is not equal to $\text{Cl}_{\text{Horn}}([\varphi \wedge \mu])$ and thus does not correspond to any Horn formula.

Assuming there exists an HPH-revision operator $\circ$ that satisfies postulates R_1, R_{3-4} as well as R_2 would immediately land us in a contradiction, since for φ and μ as in Example 6.4 we would have to conclude that $[\varphi \circ \mu] = [\varphi \wedge \mu] = \{a, b\}$, at odds with the assumption that $\varphi \circ \mu$ is a Horn formula. Note that this argument applies even if we replace R_2 with the weaker postulate R_{10} in Section 4.1, which we recall, runs as follows:

(R_{10}) If $\varphi \wedge \mu$ is consistent, then $\varphi \wedge \mu \models \varphi \circ \mu$.

For φ and μ as in Example 6.4 postulate R_{10}, in conjunction with postulate R_1, requires that $\{a, b\} \subseteq [\varphi \circ \mu] \subseteq \{a, b\}$, i.e., that $[\varphi \circ \mu] = \{a, b\}$: again, not possible.

Thus, it seems that HPH-revision operators cannot be axiomatized in a way that is analogous to $\mathcal{L}$-revision operators, at least not as long as the axiomatization is expected to include postulate R_2. Equivalently, we can state this as by saying that we cannot model an agent who, when revising a Horn formula φ by a propositional formula μ always makes the models of φ equally plausible. The reason, as we see in Example 6.4, is that, when μ is not required to be a Horn formula, the conjunction of φ and μ is not guaranteed to be a Horn formula. Thus, postulate R_2 cannot even be implemented by a well-defined HPH-revision operator. This problem extends even to the weaker postulate R_{10}, which does not explicitly require the result to be the conjunction of φ and μ, though, as we

have seen, cannot sometimes avoid it. Let us pack the morals of this discussions into one short result.

Corollary 6.1

If an HPH-revision operator satisfies postulate R_1 and R_{3-4}, then it does not satisfy either postulate R_{10} or R_2.

Since we are not prepared to sacrifice postulates R_1 and R_{3-4}, we are left with having to sacrifice postulates R_2 and R_{10}. If there is anything to salvage from postulate R_2, it is postulate R_9:

(R_9) If $\varphi \wedge \mu$ is consistent, then $\varphi \circ \mu \models \varphi \wedge \mu$.

Postulate R_9 shows promise as it can, actually, be satisfied by an HPH-revision operator, no matter the input: if $\varphi \wedge \mu$ is consistent, then, by definition, there is at least one interpretation in $[\varphi \wedge \mu]$; since singletons always correspond to some Horn formula, an HPH-revision always has something feasible it can choose, i.e., the realizability issue can be handled, in principle, by taking $[\varphi \circ \mu]$ to be a model of $\varphi \wedge \mu$. The question is whether this choice can be done in a coherent way, i.e., in a way that brings in postulates R_1 and R_{3-6}, and can be rationalized using preorders on interpretations.

The answer turns out to be yes, but under a heavy restriction of the underlying preorder. Thus, given a total, syntax insensitive $\mathcal{L}_{\mathrm{Horn}}$-assignment $\preccurlyeq$ on interpretations, we will have to impose the following property, for any interpretations w_1 and w_2:

(r_{pC}) If $w_1 \approx_\varphi w_2$, then $w_1 \subseteq w_2$ or $w_2 \subseteq w_1$.

Property r_{pC}, where 'pC' stands for *pair compliance*, says there cannot be two subset-incomparable interpretations that are equally preferred in $\leq_\varphi$. This implies that every level of a total preorder $\leq_\varphi$ in an $\mathcal{L}_{\mathrm{Horn}}$-assignment $\preccurlyeq$ on interpretations forms a chain under inclusion, i.e., that if $w_1, \ldots, w_k$ are such that $w_1 \approx_\varphi \cdots \approx_\varphi w_k$, then there exists a permutation σ such that $w_{\sigma(1)} \subseteq \cdots \subseteq w_{\sigma(k)}$. Keep in mind that the goal, here, is to ensure that an HPH-revision operator represented by an $\mathcal{L}$-assignment on interpretations is well-defined, i.e., that for any propositional formula μ, $\min_{\leq_\varphi}[\mu]$ corresponds to a Horn formula. Pair compliance turns out to guarantees this property.

Proposition 6.2

A total, syntax insensitive $\mathcal{L}$-assignment $\preccurlyeq$ on interpretations satisfies property r_{pC} if and only if $\mathrm{Cl}_{\mathrm{Horn}}(\min_{\leq_\varphi}[\mu]) = \min_{\leq_\varphi}[\mu]$, for any Horn formula φ and propositional formula μ.

Proof

("⇒") Suppose $\preccurlyeq$ satisfies property r_{pC} but there exists μ such that $\min_{\leq_\varphi}[\mu]$ is not closed under intersection. This implies that there are two interpretations w_1 and w_2 in $\min_{\leq_\varphi}[\mu]$ such that $w_1 \cap w_2 \notin \min_{\leq_\varphi}[\mu]$. Since $w_1, w_2 \in \min_{\leq_\varphi}[\mu]$, it holds that $w_1 \approx_\varphi w_2$; thus, by property r_{pC}, it follows that $w_1 \subseteq w_2$ or $w_2 \subseteq w_1$, which implies that $w_1 \cap w_2 = w_1$ or $w_1 \cap w_2 = w_2$: both cases lead to a contradiction.

("⇐") Take two interpretations w_1 and w_2 such that $w_1 \approx_\varphi w_2$, and consider $\min_{\leq_\varphi}[\varepsilon_{1,2}]$, where $\varepsilon_{1,2}$ is an $\mathcal{L}$-proxy of $\{w_1, w_2\}$, i.e., a propositional formula such that $[\varepsilon_{1,2}] = \{w_1, w_2\}$. Note that in propositional logic, $\mathcal{L}$-proxies that capture a set exactly can always be found. It follows, by our assumption, that $\min_{\leq_\varphi}[\varepsilon_{1,2}] = \{w_1, w_2\}$, which then implies that $\{w_1, w_2\}$ is closed under intersection. This, in turn, implies that $w_1 \subseteq w_2$ or $w_2 \subseteq w_1$.

The only missing piece we need is the way in which φ biases $\leq_\varphi$, but this is provided by property r_7:

(r_7) If $w_1 \in [\varphi]$ and $w_2 \notin [\varphi]$, then $w_1 <_\varphi w_2$.

We can now show that, with property r_{pC} in place, with postulate R_9 as the only sensible alternative to R_2, and with property r_7 as the semantic counterpart of postulate R_9, we can characterize HPH-revision operators along the familiar lines. If $\circ$ is an HPH-revision operator and $\preccurlyeq$ in an $\mathcal{L}_{\text{Horn}}$-assignment on interpretations, then $\circ$ *is represented by* $\preccurlyeq$, and $\preccurlyeq$ *represents* $\circ$, if, for any Horn formula φ and propositional formula μ, it holds that $[\varphi \circ \mu] = \min_{\leq_\varphi}[\mu]$.

Theorem 6.1

An HPH-revision operator $\circ$ satisfies postulates R_1, R_{3-6} and R_9 if and only if there exists a total, syntax insensitive $\mathcal{L}_{\text{Horn}}$-assignment on interpretations that satisfies properties r_7 and r_{pC} and represents $\circ$.

Proof

("⇐") If $\preccurlyeq$ is a total, syntax insensitive $\mathcal{L}_{\text{Horn}}$-assignment on interpretations that satisfies properties r_7 and r_{pC}, then we can take, as in Section 3.1, the $\preccurlyeq$-induced HPH-revision operator $\circ^{\preccurlyeq}$ by putting, for any Horn formula φ and propositional formula μ:

$$[\varphi \circ^{\preccurlyeq} \mu] \stackrel{\text{def}}{=} \min_{\leq_\varphi}[\mu].$$

Note that property r_{pC} and, in particular, Proposition 6.2, ensures that $\circ^{\preccurlyeq}$ is well-defined. Checking that $\circ^{\preccurlyeq}$ satisfies postulates R_1, R_{3-6} and R_9 follows the same lines

as in Theorems 3.1 and 4.1.

("⇒") If $\circ$ is a revision operator that satisfies postulates R_1, R_{3-6} and R_9, we can define, as in Section 3.1, the $\circ$-induced $\mathcal{L}$-assignment $\preccurlyeq^\circ$ on interpretations, for any interpretations w_1 and w_2, as follows:

$$w_1 \leq^\circ_\varphi w_2 \text{ if } w_1 \in [\varphi \circ \varepsilon_{1,2}].$$

Showing that $\leq^\circ_\varphi$ satisfies properties r_{1-4} and r_7 works the same as in Theorems 3.1 and 4.1. Property r_{pC} follows using Proposition 6.2.

Theorem 6.1 shows that the inability of HPH-revision operators to emulate postulate R_2 can be patched, to some extent, by using R_9 instead. The same cannot be said, however, for the neutrality postulate R_{NEUT}, presented first in Section 4.1:

(R_{NEUT}) $\rho(\varphi \circ \mu) \equiv \rho(\varphi) \circ \rho(\mu)$.

Postulate R_{NEUT} is innocent enough that it can normally be taken for granted, and in Chapter 4 we have seen that existing propositional distance-based revision operators satisfy it. It is worth mentioning here that this happens even though the usual postulates R_{1-6} do *not* imply postulate R_{NEUT}, even in the propositional case: an induced revision operator based on the preorder $ab <_\varphi a <_\varphi b <_\varphi \emptyset$, perfectly legal according to all postulates, suffices to make the point. However, in the propositional case postulate R_{NEUT} can be satisfied if a and b are made to be equivalent according to the preorder $\leq_\varphi$: and this is how the usual distance-based operators avoid the problem. However, in the case of HPH-revision operators, this maneuver turns out not to be possible.

Example 6.5: HPH-revision operators cannot be neutral

For the set of atoms $A = \{a, b\}$, consider the Horn formula $\varphi = a \wedge b$, the propositional formula $\mu = a \leftrightarrow \neg b$ and an HPH-revision operator $\circ$ that satisfies postulates R_1, R_3 and R_4. By postulates R_1 and R_3, we have that $[\varphi \circ \mu]$ is a non-empty subset of $[\mu] = \{a, b\}$. Since $\varphi \circ \mu$ is, by definition, a Horn formula, it has to be the case either that $[\varphi \circ \mu] = \{a\}$, or $[\varphi \circ \mu] = \{b\}$, which implies that either $\varphi \circ \mu \equiv a \wedge \neg b$ or $\varphi \circ \mu \equiv \neg a \wedge b$.

Without loss of generality, assume that $\varphi \circ \mu \equiv a \wedge \neg b$, and take a renaming ρ such that $\rho(a) = b$ and $\rho(b) = a$. It follows, then, that $\rho(\varphi \circ \mu) = b \wedge \neg a$. At the same time, we have that that $\rho(\varphi) = b \wedge a \equiv \varphi$ and $\rho(\mu) = b \leftrightarrow \neg a \equiv \mu$, which by postulate R_4 implies that $\rho(\varphi) \circ \rho(\mu) \equiv \varphi \circ \mu \equiv a \wedge \neg b$. Thus, in this case we have that $\rho(\varphi \circ \mu) \not\equiv \rho(\varphi) \circ \rho(\mu)$, a result inconsistent with postulate R_{NEUT}.

Example 6.5 points to a fundamental contradiction at the heart of HPH-revision operators supposed to satisfy the, arguably undisputable, postulates R_1 and R_{3-4}: they cannot

be neutral. Intuitively, this occurs because revision by a formula $\mu = a \leftrightarrow \neg b$ must return a consistent result that implies μ and is a Horn formula. In other words, such an operator must effectively choose exactly one of the interpretations a and b: this leads to a clash with the neutrality postulate $\mathrm{R}_{\mathsf{NEUT}}$, which tries to prevent this sort of preferential behavior. Example 6.5 translates into the following result.

Corollary 6.2

If an HPH-revision operator satisfies postulates R_1 and R_{3-4}, then it does not satisfy postulate $\mathrm{R}_{\mathsf{NEUT}}$.

The move to be explicit about neutrality and to split the standard postulate R_2 into two distinct properties (postulates R_9 and R_{10}), either of which can be turned off, finds additional justification here: we can see now that properties taken for granted in the propositional case break down when restricting the language, and a thorough analysis of what are rational, or desirable, properties for revision must take this into account.

6.3 Horn revision by Horn formulas

In this section we look at revision of Horn formulas when both inputs, as well as the output are in the Horn fragment. This part about total preorders mostly recapitulates existing results [Delgrande and Peppas, 2015, Delgrande et al., 2018], but the main storyline will be important for the remaining parts.

An HHH-*revision operator* $\circ$ is a function $\circ\colon \mathcal{L}_{\mathrm{Horn}} \times \mathcal{L}_{\mathrm{Horn}} \to \mathcal{L}_{\mathrm{Horn}}$, taking as input two Horn formulas, typically denoted by φ and μ and referred to as the prior and new information, respectively, and returning a Horn formula, typically denoted by $\varphi \circ \mu$ and referred to as the posterior information.

The postulates we want to make use of in this section are the standard revision postulates R_{1-8} presented in Section 3.1, but particularized to Horn formulas. These postulates, we will say, apply to any Horn formulas φ, φ_1, φ_2, μ, μ_1 and μ_2:

(R_1) $\varphi \circ \mu \models \mu$.

(R_2) If $\varphi \wedge \mu$ is consistent, then $\varphi \circ \mu \equiv \varphi \wedge \mu$.

(R_3) If μ is consistent, then $\varphi \circ \mu$ is consistent.

(R_4) If $\varphi_1 \equiv \varphi_2$ and $\mu_1 \equiv \mu_2$, then $\varphi_1 \circ \mu_1 \equiv \varphi_2 \circ \mu_2$.

(R_5) $(\varphi \circ \mu_1) \wedge \mu_2 \models \varphi \circ (\mu_1 \wedge \mu_2)$.

(R_6) If $(\varphi \circ \mu_1) \wedge \mu_2$ is consistent, then $\varphi \circ (\mu_1 \wedge \mu_2) \models (\varphi \circ \mu_1) \wedge \mu_2$.

(R_7) If $\varphi \circ \mu_1 \models \mu_2$ and $\varphi \circ \mu_2 \models \mu_1$, then $\varphi \circ \mu_1 \equiv \varphi \circ \mu_2$.

(R$_8$) If $\mu \equiv \mu_1 \vee \mu_2$, then $(\varphi \circ \mu_1) \wedge (\varphi \circ \mu_2) \models \varphi \circ \mu$.

The intuitions guiding postulates R$_{1-8}$ are the same whether they apply to Horn formulas or to propositional formulas, and can be consulted in Section 3.1. As for revision, postulates R$_7$ and R$_8$ are weaker than postulate R$_6$, and we will think of them as alternatives to R$_6$. Note that, since the Horn fragment is closed under conjunction, postulate R$_2$ (and every other postulate that uses the conjunction of two formulas) can be used here. Note, also, that postulate R$_8$ applies here only to Horn formulas μ, μ_1 and μ_2 such that $\mu \equiv \mu_1 \vee \mu_2$, i.e., to Horn formulas μ_1 and μ_2 such that their disjunction is also a Horn formula. Since the disjunction of two Horn formulas is not guaranteed to be a Horn formula (see Example 6.2), this effectively amounts to restricting postulate R$_8$ to only a subset of the formulas in the language we are working in.

On the semantic side, we will work with an $\mathcal{L}_{\text{Horn}}$-assignment $\preccurlyeq$ on interpretations, i.e., a function that maps every Horn formula φ to a ranking $\leq_\varphi$ on interpretations. The rankings in an $\mathcal{L}_{\text{Horn}}$-assignment on interpretations expected to satisfy some subset of the following properties, for any Horn formulas φ, φ_1, φ_2 and interpretations w_1, w_2 and w_3:

(r$_1$) $w \leq_\varphi w$.

(r$_2$) If $w_1 \leq_\varphi w_2$ and $w_2 \leq_\varphi w_3$, then $w_1 \leq_\varphi w_3$.

(r$_3$) $w_1 \leq_\varphi w_2$ or $w_2 \leq_\varphi w_1$.

(r$_4$) If $\varphi_1 \equiv \varphi_2$, then $w_1 \leq_{\varphi_1} w_2$, then if $w_1 \leq_{\varphi_2} w_2$.

(r$_5$) If $w_1, w_2 \in [\varphi]$, then $w_1 \approx_\varphi w_2$.

(r$_6$) If $w_1, w_2 \in [\varphi]$, then $w_1 \not<_\varphi w_2$ and $w_2 \not<_\varphi w_1$.

(r$_7$) If $w_1 \in [\varphi]$ and $w_2 \notin [\varphi]$, then $w_1 <_\varphi w_2$.

Properties r$_{1-7}$ are familiar from Section 3.1, and they amount to the same expectation: that $\leq_\varphi$ is a preorder, partial or total, that makes the models of φ the minimal elements in $\leq_\varphi$. The next notions are inherited from the propositional case, and we rehearse them here only in the spirit of completeness. An $\mathcal{L}_{\text{Horn}}$-assignment $\preccurlyeq$ on interpretations is *partial* if it satisfies properties r$_{1-2}$, *total* if it satisfies properties r$_{1-3}$, *syntax insensitive* if it satisfies property r$_4$ and *r-faithful* if it satisfies properties r$_{6-7}$. The usual caveat applies, that if $\leq_\varphi$ is total and satisfies property r$_6$, then it also satisfies property r$_5$. If $\circ$ is an HHH-revision operator and $\preccurlyeq$ is an $\mathcal{L}_{\text{Horn}}$-assignment on interpretations, then $\preccurlyeq$ *represents* $\circ$ (and $\circ$ *is represented by* $\preccurlyeq$) if, for any Horn formulas φ and μ, it holds that $[\varphi \circ \mu] = \min_{\leq_\varphi}[\mu]$. Given an $\mathcal{L}_{\text{Horn}}$-assignment $\preccurlyeq$ on interpretations, the $\preccurlyeq$*-induced* $\mathcal{L}$*-revision operator* $\circ^{\preccurlyeq}$ is defined, for any Horn formulas φ and μ, by taking:

$$[\varphi \circ^{\preccurlyeq} \mu] \stackrel{\text{def}}{=} \min_{\leq_\varphi}[\mu].$$

Note that we have defined $\circ^{\preccurlyeq}$ as an $\mathcal{L}$-revision operator, i.e., an operator that returns propositional formulas, not necessarily Horn formula: this is a precaution against problems that can arise from $\min_{\leq_\varphi}[\mu]$ not being Horn-closed. It is clear that $\min_{\leq_\varphi}[\mu]$ represents some propositional formula, but whether it represents a Horn formula is not obvious at this point. And indeed, we will see that this precaution is warranted.

The next step involves reconstructing the plausibility relations from the information provided by the revision operator: as per usual we will compare interpretations using their proxy formulas: however, since we are only allowed to use Horn formulas, proxy formulas cannot be as precise as they are when working in propositional logic. Thus, if w_1 and w_2 are interpretations such that $w_1 \not\subseteq w_2$ and $w_2 \not\subseteq w_1$, then the models of a $\mathcal{L}_{\text{Horn}}$-proxy formula $\varepsilon_{1,2}$ are $[\varepsilon_{1,2}] = \{w_1, w_2, w_1 \cap w_2\}$. For the next definition, keep in mind that $\varepsilon_{1,2}$ refers to the $\mathcal{L}_{\text{Horn}}$-proxy of two interpretations w_1 and w_2.

Given an HHH-revision operator $\circ$ and a Horn formula φ, the *exhaustive $\circ$-revealed plausibility relation* $\leq_\varphi^{\text{exh}}$ and the *exclusive $\circ$-revealed plausibility relation* $\leq_\varphi^{\text{exc}}$ are defined, for any interpretations w_1 and w_2, respectively, as:

$$w_1 \leq_\varphi^{\text{exh}} w_2 \text{ if } w_1 \in [\varphi \circ \varepsilon_{1,2}],$$
$$w_1 \leq_\varphi^{\text{exc}} w_2 \text{ if either } w_1 = w_2, \text{or } w_1 \in [\varphi \circ \varepsilon_{1,2}] \text{ and } w_2 \notin [\varphi \circ \varepsilon_{1,2}].$$

The *exhaustive $\mathcal{L}_{\text{Horn}}$-revealed assignment* $\preccurlyeq^{\text{exh}}$ and *exclusive $\mathcal{L}_{\text{Horn}}$-revealed assignment* $\preccurlyeq^{\text{exc}}$ are obtained by taking $\preccurlyeq^{\text{exh}}(\varphi) = \leq_\varphi^{\text{exh}}$ and $\preccurlyeq^{\text{exc}}(\varphi) = \leq_\varphi^{\text{exc}}$, for any Horn formula φ.

Since we generally assume that the revision operators we work with satisfy postulates R_1 and R_{3-4}, it follows that both the exhaustive and exclusive revealed assignments are reflexive out of the box. For the exclusive revealed assignment, we additionally have that preorders $\leq_\varphi^{\text{exc}}$ are strict for any distinct interpretations w_1 and w_2, i.e., if $w_1 \neq w_2$ then $w_1 <_\varphi^{\text{exc}} w_2$.

However, unlike in propositional logic, the exhaustive revealed relation is not guaranteed to be complete anymore: indeed, if $[\varphi \circ \varepsilon_{1,2}] = \{w_1 \cap w_2\}$, then w_1 and w_2 are incomparable with respect to $\leq_\varphi^{\text{exh}}$. What is more, the exhaustive revealed relation is not even guaranteed to be transitive.

Example 6.6: $\leq_\varphi^{\text{exh}}$ might not be transitive

Consider an HHH-revision operator $\circ$ such that, for the set of atoms $A = \{a, b\}$, delivers the following results: $[\varphi \circ \varepsilon_{a,ab}] = \{a\}$, $[\varphi \circ \varepsilon_{ab,b}] = \{ab\}$ and $[\varphi \circ \varepsilon_{a,b}] = \{\emptyset\}$. From this we infer that $a <_\varphi^{\text{exh}} ab$, $ab <_\varphi^{\text{exh}} b$, $\emptyset <_\varphi^{\text{exh}} a$ and $\emptyset <_\varphi^{\text{exh}} b$. These comparisons are depicted in Figure 6.1. In propositional logic we would be able to use postulates R_{5-6} to infer from the result for $\varphi \circ \varepsilon_{a,ab}$ and for $\varphi \circ \varepsilon_{ab,b}$ that a choice over interpretations a, ab and b has to select a; based on this, we would infer that the choice over a and b has to select a as well, meaning that a is considered better than b. However, in the Horn fragment there is no way of enforcing choice over a, ab

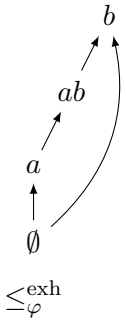


Figure 6.1: The exhaustive $\circ$-revealed $\mathcal{L}_{\text{Horn}}$-assignment is not guaranteed to be transitive.

and b, since $\{a, b, ab\}$ is not Horn-closed. This makes it possible for the result to be $\emptyset$ when revising by $\varepsilon_{a,b}$, leaving a and b incomparable in $\leq_\varphi^{\text{exh}}$.

Example 6.6 illustrates one of the consequences of restricting the language: since there are certain configurations of outcomes the agent never gets to see, the revision operator becomes less precise at identifying the relationship between certain outcomes: if $w_1 \subseteq w_2$ or $w_2 \subseteq w_1$, then $\{w_1, w_2\}$ is Horn-closed and $\varphi \circ \varepsilon_{1,2}$ behaves as in propositional logic; but if w_1 and w_2 are subset-incomparable, then there is no way to make a comparison only between the two of them, and the revealed relation may feature patches where the revision operators has nothing informative to say.

Nevertheless, a quick glance at Example 6.6 suggests an easy fix: it is straightforward to see that the transitive closure of $\leq_\varphi^{\text{exh}}$, as depicted in Figure 6.1, is an ordering extension of $\leq_\varphi^{\text{exh}}$, as defined in Section 2.4, i.e., a relation that preserves all the comparisons in $\leq_\varphi^{\text{exh}}$, including the strict ones. Even if the revision operator does not explicitly state that a is better than b, given the prior information φ, we can still infer this from the intermediary comparisons of $a <_\varphi^{\text{exh}} ab$ and $ab <_\varphi^{\text{exh}} b$. Importantly, adding the comparison $a <_\varphi^{\text{exh}} b$ to $\leq_\varphi^{\text{exh}}$ does not misrepresent $\circ$: the augmented relation underlies the same revision operator. This raises the hope of a more general strategy, and we will soon see the conditions under which this strategy is successful.

So far, so good. The problem, as has been already documented [Delgrande and Peppas, 2015, Delgrande et al., 2018], is that these elements alone are not enough to deliver a representation result. A first indication that more needs to be done is the fact that assignments based on standard preorders (either total or partial, r-faithful or not) cannot be used to induce HHH-revision operators.

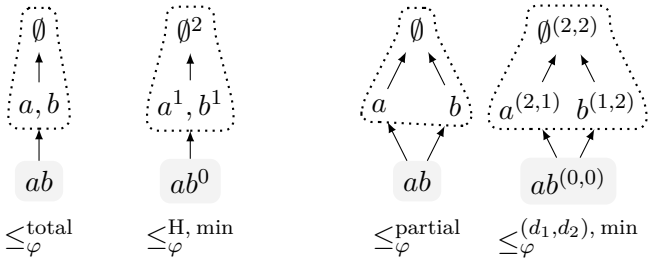


Figure 6.2: Preorders $\leq_\varphi^{\text{total}}$ and $\leq_\varphi^{\text{partial}}$ for $\varphi = a \wedge b \wedge \neg c$ and $\mu = \neg(a \wedge b) \wedge \neg c$. Models of φ are shaded in gray, models of μ are surrounded by the dotted line. Neither of the preorders $\leq_\varphi^{\text{total}}$ and $\leq_\varphi^{\text{partial}}$ delivers a Horn-closed set of interpretations, and thus cannot be used to model HHH-revision operators. These preorders coincide with $\leq_\varphi^{\text{H, min}}$ and $\leq_\varphi^{(d_1,d_2),\,\text{min}}$, as presented in Example 3.6.

Example 6.7: Preorders that do not induce an HHH-revision operator

For the set of atoms $A = \{a, b, c\}$, consider formulas $\varphi = a \wedge b \wedge \neg c$ and $\mu = \neg(a \wedge b) \wedge \neg c$. We have that $[\varphi] = \{ab\}$ and $[\mu] = \{\emptyset, a, b\}$. Both φ and μ are Horn formulas, and are thus valid inputs to an HHH-revision operator. Consider, however, a total assignment $\preccurlyeq^{\text{total}}$ and a partial assignment $\preccurlyeq^{\text{partial}}$ that assigns to φ the preorders $\leq_\varphi^{\text{total}}$ and $\leq_\varphi^{\text{partial}}$, respectively, both depicted in Figure 6.2. We obtain that:

$$\begin{aligned}\min_{\leq_\varphi^{\text{total}}}[\mu] &= \min_{\leq_\varphi^{\text{partial}}}[\mu] \\ &= \{a, b\},\end{aligned}$$

and hence $[\varphi \circ^{\text{total}} \mu] = [\varphi \circ^{\text{partial}} \mu] = \{a, b\}$. Since $\text{Cl}_{\text{Horn}}(\{a, b\}) = \{\emptyset, a, b\} \neq \{a, b\}$, it follows that $[\varphi \circ^{\text{total}} \mu]$ and $[\varphi \circ^{\text{partial}} \mu]$ cannot be represented as a Horn formula, i.e., the $\preccurlyeq^{\text{total}}$-induced and $\preccurlyeq^{\text{partial}}$-induced revision operators do not work as HHH-revision operators.

This finding is significant, because $\leq_\varphi^{\text{total}}$ coincides on the models of φ and μ with the preorder generated by many of the distance-based assignments we looked at in Section 3.1 (the set of atoms is considered to coincide with A in that example), namely with $\leq_\varphi^{\text{H, min}}$, $\leq_\varphi^{\text{H, leximin}}$, $\leq_\varphi^{\text{H, max}}$, $\leq_\varphi^{\text{H, leximax}}$ and $\circ_\varphi^{\text{H, sum}}$. Also, $\leq_\varphi^{\text{partial}}$ coincides on the models of φ and μ with the partial preorder $\leq_\varphi^{(d_1,d_2),\,\text{min}}$ presented in Example 3.6.

Example 6.7 shows that, like in Section 6.2, there are certain preorders bound to deliver results that cannot be recast as Horn formulas. Unfortunately, these preorders appear in the distance-based assignments we have introduced in Section 3.1 for $\mathcal{L}$-revision operators. The upshot, then, is that none of these assignments can be repurposed for HHH-revision.

The problem highlighted by Example 6.7 is that properties r_{1-7}, by themselves, make an $\mathcal{L}_{\text{Horn}}$-assignment $\preccurlyeq$ on interpretations ill-equipped to induce a well-defined HHH-revision

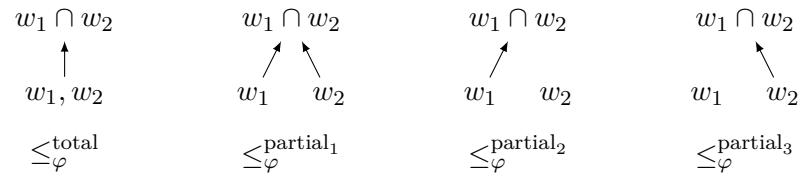


Figure 6.3: Preorders, total and partial, that deliver results that are not Horn-closed when revising by a Horn formula μ, with $[\mu] = \{w_1, w_2, w_1 \cap w_2\}$. It is assumed that $w_1 \not\subseteq w_2$ and $w_2 \not\subseteq w_1$, such that $w_1 \cap w_2$ is an interpretation distinct from both w_1 and w_2. The goal of property r_{HC} is to prevent the occurrence of these configurations.

operator. The solution, as for HPH-revision operators, is to rein in the preorders we are looking at. The purpose, here, is to find a restriction on $\mathcal{L}_{\text{Horn}}$-assignments that ensures they deliver Horn-closed results when the new information is a Horn formula. This is done via the following property, intended to hold for any Horn formula φ and interpretations w_1 and w_2:

(r_{HC}) For any Horn formula μ, it holds that $\min_{\leq_\varphi}[\mu]$ if Horn-closed.

Property r_{HC}, where 'HC' stands for *Horn compliance* [Delgrande and Peppas, 2015, Delgrande et al., 2018], guarantees that the $\leq_\varphi$-minimal elements of $[\mu]$, when μ is a Horn formula, can be represented by a Horn formula. Property m_{HC} works for both total and partial preorders, and its role, essentially, is to rule out situations such as the ones in Figure 6.3, where revision by a Horn formula yields a set of interpretations that cannot be expressed as a Horn formula. An $\mathcal{L}_{\text{Horn}}$-assignment $\preccurlyeq$ on interpretations is *Horn compliant* if it satisfies property r_{HC}. It is straightforward to see that property r_{HC} is a condition that is both necessary and sufficient for a propositional revision operator to function as an HHH-revision operator.

With the expressibility issue fixed, the next step is to look at the effect of postulates R_{1-6}, or R_{1-5} and R_{7-8}, and connect them to properties r_{1-7}. However, another problem rears its head: it turns out that restricted to Horn formulas, the revision postulates end up saying less than their propositional counterparts, to the point where they can now induce unwanted assignments. To understand this issue, it is best to look at total and partial preorders separately.

Total preorders

Total preorders, we know from Section 3.1, go hand in hand with postulates R_{1-6}, but we are beginning to see that for the Horn fragment these elements may not be enough. As has been observed, one outstanding problem is the presence of non-transitive cycles in assignments that can represents HHH-revision operators that satisfy postulates R_{1-6}.

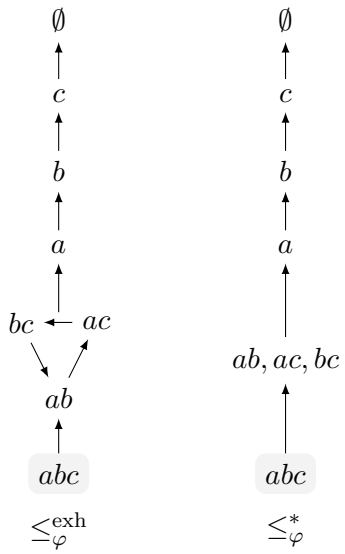


Figure 6.4: The exhaustive revealed ranking $\leq_\varphi^{\text{exh}}$ for an HHH-revision operator $\circ$ that satisfies postulates R_{1-5} and R_{7-8} and $\varphi = a \wedge b \wedge c$, together with the ranking $\leq_\varphi^*$ obtained as the transitive closure of $\leq_\varphi^{\text{exh}}$. The ranking $\leq_\varphi^{\text{exh}}$ contains a non-transitive cycle between ab, ac and bc, and behaves like a total preorder otherwise. The non-transitive cycle goes undetected when $\circ$ processes only Horn formulas. The transitive closure $\leq_\varphi^*$ of $\leq_\varphi^{\text{exh}}$ fixes the non-transitivity issue but does not represent the revision operator $\circ$ anymore.

Example 6.8: [Delgrande and Peppas, 2015, Delgrande et al., 2018]

For the set of atoms $A = \{a, b, c\}$ and the formula $\varphi = a \wedge b \wedge c$, consider an HHH-revision operator $\circ$ that induces the exhaustive revealed plausibility relation $\leq_\varphi^{\text{exh}}$ in Figure 6.4. In other words, it holds that $[\varphi \circ \varepsilon_{ab,ac}] = \{ab\}$, $[\varphi \circ \varepsilon_{ac,bc}] = \{ac\}$, $[\varphi \circ \varepsilon_{bc,ab}] = \{bc\}$, and so on: the results of the revision operator on all the possible inputs can be read off from Figure 6.4.

The significant detail about the revealed relation $\leq_\varphi^{\text{exh}}$ is that it behaves like a total preorder everywhere except on ab, ac and bc, which are fixed into a non-transitive cycle. In other words, $\circ$ decides that $ab <_\varphi^{\text{exh}} ac <_\varphi^{\text{exh}} bc <_\varphi^{\text{exh}} ab$, which implies that $\leq_\varphi$ does not satisfy property r_3 on ab, ac and bc.

We would expect that the failure of $\leq_\varphi^{\text{exh}}$ to satisfy property r_3 translates into $\circ$ not satisfying some of the revision postulates: and if we were working in propositional logic, this would indeed be the result. In propositional logic we can always revise by a propositional formula that has exactly ab, ac and bc as its models; revision with this formula together with postulate R_3 implies that $\min_{\leq_\varphi^{\text{exh}}}\{ab, ac, bc\}$ has to be non-empty, contrary to the present situation: in the framework of propositional logic

the regular postulates R_{1-6} make a revision operator such as the operator $\circ$ specified here impossible.

Interestingly, in the Horn fragment the operator $\circ$ turns out to be a perfectly legal HHH-revision operator: it can be checked that $\circ$ satisfies postulates R_{1-6} [Delgrande and Peppas, 2015, Delgrande et al., 2018]. The reason why the cycle manages to slip through undetected is that in the Horn fragment there is no formula that has exactly ab, ac and bc as its models, since $\{ab, ac, bc\}$ is not Horn-closed. The closest we can come to this is by using an $\mathcal{L}_{\mathrm{Horn}}$-proxy $\varepsilon_{ab,ac,bc}$ of $\{ab, ac, bc\}$, but $[\varepsilon_{ab,ac,bc}] = \mathrm{Cl}_{\mathrm{Horn}}(\{ab, ac, bc\}) = \{\emptyset, a, b, c, ab, ac, bc\}$, and simply asking that $\min^{\mathrm{exh}}_{\leq_\varphi}[\varepsilon_{ab,ac,bc}]$ is non-empty does not prevent the cycle.

A tentative fix for this situation is to replace $\leq^{\mathrm{exh}}_\varphi$, as in Example 6.6 with its transitive closure $\leq^*_\varphi$, also depicted in Figure 6.4. This has the effect of flattening the cycle by introducing indifference between ab, bc and ac. The downside of this move however, is that $\leq^*_\varphi$ does not preserve the information provided by $\circ$: we can see this by looking at the revision operator $\circ^*$ induced by the preorder $\leq^*_\varphi$, and comparing it to $\circ$: we have that $[\varphi \circ \varepsilon_{ab,ac}] = \{ab\}$, i.e., $\min_{\leq^{\mathrm{exh}}_\varphi}\{ab, ac, a\} = \{ab\}$, whereas $\min_{\leq^*_\varphi}\{ab, ac, a\} = \{ab, ac\}$, i.e., $[\varphi \circ^* \varepsilon_{ab,ac}] = \{ab, ac\}$. The cycle-free transitive closure of $\leq^{\mathrm{exh}}_\varphi$ represents a different revision operation, one that does not even return a Horn formula!

The moral here is that the preorder $\leq^{\mathrm{exh}}_\varphi$ depicted in Figure 6.4 cannot be extended to a total preorder while still remaining faithful to the revision operator $\circ$: eliminating the cycle between ab, bc and ac would lead to $\leq^{\mathrm{exh}}_\varphi$ misrepresenting $\circ$. The cycle, here, is unavoidable.

In Example 6.6 we encountered an exhaustive HHH-revision operator that induced a non-transitive ranking on outcomes, but this ranking could be made transitive by filling in the gaps with the comparisons inferred by transitivity. Example 6.4 shows that there are exhaustive HHH-revision operators inducing rankings that are not only non-transitive, but that cannot even be made transitive: postulates R_{1-6}, the most demanding revision postulates we have, not only fail to notice the cycle between ab, ac and bc, but allow the agent to revise in a way that makes the cycle compulsory. This is a direct result of the Horn fragment's inability to capture certain sets of interpretations: whereas in propositional logic we would be able to leverage postulates R_{1-6} to make sure that the revealed exhaustive ranking is transitive, in the Horn fragment this move is not possible.

One way to deal with this situation is to make sure that an exhaustive HHH-revision operator does not revise in a way that paints it into a non-transitive corner, and it is here that the literature on rational choice proves useful, as it suggests a tool of proven efficacy: Suzumura consistency. Recall Theorem 2.1, saying that Suzumura consistency is both a necessary and sufficient condition for a binary relation to have an extension that is a total preorder. In the context of revision, we can formulate Suzumura consistency as

a property that applies to preorders in an assignment $\preccurlyeq$ on interpretations. Thus, for any Horn formula φ and interpretations $w_1, \ldots, w_n$, the property is as follows:

($\mathrm{r_{SC}}$) If $w_1 \leq_\varphi \cdots \leq_\varphi w_n$, then $w_n \not<_\varphi w_1$.

Property $\mathrm{r_{SC}}$, with SC standing for *Suzumura consistency*, has a natural reading: if w_1 is at least as good as w_2 according to $\leq_\varphi$, w_2 is at least as good as w_3, and so on, all the way to w_n, then the betterness of w_1 should propagate down the line, i.e., the last outcome in this sequence cannot be strictly better than w_1. Property $\mathrm{r_{SC}}$ is, of course, implied by transitivity, i.e., by property $\mathrm{r_3}$, but if property $\mathrm{r_3}$ cannot be enforced then $\mathrm{r_{SC}}$ is the safest bet, as long as $\mathrm{r_{SC}}$ itself can be enforced. The way this is done is by supplementing the standard set of postulates with a special postulate, tailored specifically for property $\mathrm{r_{SC}}$. We will formulate the postulate using the notion of an $\mathcal{L}_{\mathrm{Horn}}$-proxy of a set $\mathcal{W}$ of interpretations, introduced in Section 6.1: recall, this is a Horn formula $\varepsilon_{\mathcal{W}}$ such that $[\varepsilon_{\mathcal{W}}] = \mathrm{Cl_{Horn}}(\mathcal{W})$. We will apply this notion to singletons $\{w_i\}$ and pairs $\{w_i, w_j\}$ of interpretations, in which case we write ε_i and $\varepsilon_{i,j}$ instead of ε_{w_i} and ε_{w_i,w_j}, respectively. Since singleton sets of interpretations are Horn-closed, it holds that $[\varepsilon_i] = \mathrm{Cl_{Horn}}(\{w_i\}) = \{w_i\}$. Pairs of interpretations are not necessarily Horn-closed, in which case $[\varepsilon_{i,j}]$ may contains the additional interpretation $w_i \cap w_j$.

The additional postulate, or, more precisely, postulate schema, is intended to work for the exhaustive revealed assignment, and meant to apply for any Horn formula φ, integer n, interpretations $w_1, \ldots, w_n$ and their associate $\mathcal{L}_{\mathrm{Horn}}$-proxy formulas:

($\mathrm{R_{SC}}$) If $(\varphi \circ \varepsilon_{1,2}) \wedge \varepsilon_1$ is consistent, $\ldots$, $(\varphi \circ \varepsilon_{n-1,n}) \wedge \varepsilon_{n-1}$ is consistent, then it does not hold that both $(\varphi \circ \varepsilon_{n,1}) \wedge \varepsilon_n$ is consistent and that $(\varphi \circ \varepsilon_{n,1}) \wedge \varepsilon_1$ is inconsistent.

Postulate $\mathrm{R_{SC}}$ expresses the same idea as property $\mathrm{r_{SC}}$, but using formulas and the revision operator instead of interpretations and the preorder. Note that this connection applies only if the preorder is part of the exhaustive $\circ$-revealed assignment.

Theorem 6.2

If $\circ$ is an HHH-revision operator and $\preccurlyeq^{\mathrm{exh}}$ is the $\circ$-revealed exhaustive assignment, then $\circ$ satisfies postulate $\mathrm{R_{SC}}$ if and only if $\preccurlyeq^{\mathrm{exh}}$ satisfies property $\mathrm{r_{SC}}$.

Proof

("$\Rightarrow$") Assume $\circ$ satisfies postulate $\mathrm{R_{SC}}$ and take interpretations $w_1, \ldots, w_n$ such that $w_1 \leq_\varphi^{\mathrm{exh}} \cdots \leq_\varphi^{\mathrm{exh}} w_n$, and suppose $w_n <_\varphi^{\mathrm{exh}} w_1$. It follows, first, that $w_1 \in [(\varphi \circ \varepsilon_{1,2}) \wedge \varepsilon_1]$, $\ldots$, $w_{n-1} \in [(\varphi \circ \varepsilon_{n-1,n}) \wedge \varepsilon_{n-1}]$, which shows that the precondition of postulate $\mathrm{R_{SC}}$ is satisfied. The assumption that $w_n <_\varphi^{\mathrm{exh}} w_1$ implies that $[(\varphi \circ \varepsilon_{n,1}) \wedge \varepsilon_n] = \{w_n\}$,

i.e., that $(\varphi \circ \varepsilon_{n,1}) \wedge \varepsilon_n$ is consistent and $(\varphi \circ \varepsilon_{n,1}) \wedge \varepsilon_1$ is inconsistent, which is a contradiction.

("$\Leftarrow$") Suppose $(\varphi \circ \varepsilon_{1,2}) \wedge \varepsilon_1$ is consistent, ..., $(\varphi \circ \varepsilon_{n-1,n}) \wedge \varepsilon_{n-1}$ is consistent, and, in addition, that $(\varphi \circ \varepsilon_{n,1}) \wedge \varepsilon_n$ is consistent and that $(\varphi \circ \varepsilon_{n,1}) \wedge \varepsilon_1$ is inconsistent. It follows from this that $w_1 \leq_\varphi^{\text{exh}} \cdots \leq_\varphi^{\text{exh}} w_n$, and $w_n <_\varphi^{\text{exh}} w_1$; assuming that $\preccurlyeq^{\text{exh}}$ satisfies property r_{SC}, this leads to a contradiction.

It must be mentioned that R_{SC} is no more than a rewriting of the acyclicity postulate that has already been shown to work alongside property r_{SC} in axiomatizing HHH-revision operators, while following from postulates R_{1-6} in propositional logic [Delgrande and Peppas, 2015, Delgrande et al., 2018]. What the current discussion adds is only some context from rational choice theory, which allows us to see the existing results in a different light. Thus, Theorem 6.2 shows that adding postulate R_{SC} guarantees that the preorders in the exhaustive revealed assignment are Suzumura consistent, and therefore, by Theorem 2.1, can be extended to a total preorder: postulate R_{SC}, in other words, eliminates the cycles such as the one in Example 6.8, and is exactly the property we need to make sure that the exhaustive revealed preorder can still be extended to a total preorder. This is important for the prospects of a representation theorem: recall that our goal is to show that HHH-revision operators satisfying postulates R_{1-6} can be represented using total assignments on interpretations: with postulate R_{SC}, the exhaustive revealed assignment can be seen to be a promising candidate, since it manages to represent $\circ$ and admits of ordering extensions. Significantly, existing work [Delgrande and Peppas, 2015] shows how to construct such an ordering. This procedure starts by extending $\leq_\varphi^{\text{exh}}$ to its transitive closure, which, by design, guarantees transitivity. However, the transitive closure is still not guaranteed to be total, and the construction further contains a way of resolving incomparabilities in a way that does not disturb the minimal models of any Horn formula μ. The last step is important, since it ensures that the extension still manages to represent the operator $\circ$. This construction, therefore, is at the heart of the following representation theorem for HHH-revision operators.

Theorem 6.3 ([Delgrande and Peppas, 2015])

An HHH-revision operator $\circ$ satisfies postulates R_{1-6} and R_{SC} if and only if there exists an $\mathcal{L}_{\text{Horn}}$-assignment $\preccurlyeq$ on interpretations that satisfies properties r_{1-7} and r_{HC} (i.e., is total, syntax independent, r-faithful and Horn compliant) and that represents the operator $\circ$.

What we can add here to this result is the observation that, indeed, *any* ordering extension of the $\circ$-revealed exhaustive assignment is guaranteed to work. Let us unpack this statement by taking stock of where we are at this point. Recall that, if φ is a Horn formula, the exhaustive revealed ranking $\leq_\varphi^{\text{exh}}$ is not guaranteed to be total, even with postulates R_{1-6} in place: if w_1 and w_2 are subset-incomparable interpretations, then

they can end up being incomparable in $\leq_\varphi^{\text{exh}}$ if $[\varphi \circ \varepsilon_{1,2}] = \{w_1 \cap w_2\}$, where $\varepsilon_{1,2}$ is an $\mathcal{L}_{\text{Horn}}$-proxy of $\{w_1, w_2\}$, i.e., a Horn formula such that $[\varepsilon_{1,2}] = \{w_1, w_2, w_1 \cap w_2\}$. What is more, subset-incomparable intepretations are the *only* pairs of interpretations that can end up being incomparable in $\leq_\varphi^{\text{exh}}$: if $w_1 \subseteq w_2$ or $w_2 \subseteq w_1$, then the HHH-revision operator $\circ$ can compare w_1 and w_2 directly, and it holds that $w_1 \leq_\varphi^{\text{exh}} w_2$ or $w_2 \leq_\varphi^{\text{exh}} w_1$. At the same time, because of postulate R_{SC}, $\leq_\varphi^{\text{exh}}$ is guaranteed to satisfy property r_{SC}. Thus, by Theorem 2.1, this means that there exists an ordering extension of $\leq_\varphi^{\text{exh}}$, i.e., a total preorder that preserves all comparisons in $\leq_\varphi^{\text{exh}}$, including, significantly, the strict ones. What we can show now is that any such ordering extension leaves the minimal elements of any Horn formula μ unchanged.

Proposition 6.3

If $\circ$ is an HHH-revision operator that satisfies postulates R_{1-6} and R_{SC} and φ is a Horn formula, then any ordering extension $\leq_\varphi^{\text{exh}*}$ of the ranking $\leq_\varphi^{\text{exh}}$ assigned to φ by the $\circ$-revealed exhaustive assignment $\preccurlyeq^{\text{exh}}$ is such that $\min_{\leq_\varphi^{\text{exh}}}[\mu] = \min_{\leq_\varphi^{\text{exh}*}}[\mu]$, for any Horn formula μ.

Proof

("$\subseteq$") Take $w_1 \in \min_{\leq_\varphi^{\text{exh}}}[\mu]$ and suppose that $w_1 \notin \min_{\leq_\varphi^{\text{exh}*}}[\mu]$. This means that there exists an interpretation $w_2 \in [\mu]$ such that $w_2 <_\varphi^{\text{exh}*} w_1$. Let us reflect, now on what the situation of w_1 and w_2 can be in $\leq_\varphi^{\text{exh}}$. Since $\leq_\varphi^{\text{exh}*}$ is an ordering extension of $\leq_\varphi^{\text{exh}}$, it cannot be the case that $w_1 \leq_\varphi^{\text{exh}} w_2$, since this would imply that $w_1 \leq_\varphi^{\text{exh}*} w_2$. The only possibilities, therefore, are either that $w_2 <_\varphi^{\text{exh}} w_1$, or w_1 and w_2 are incomparable in $\leq_\varphi^{\text{exh}}$. The case where $w_2 <_\varphi^{\text{exh}} w_1$ contradicts the assumption that $w_1 \in \min_{\leq_\varphi^{\text{exh}}}[\mu]$. But if w_1 and w_2 are incomparable in $\leq_\varphi^{\text{exh}}$, then it must be the case that w_1 and w_2 are subset incomparable and $[\varphi \circ \varepsilon_{1,2}] = \{w_1 \cap w_2\}$. Using postulates R_{1-6} we now show, as we have done before, that it also holds that $\varphi \circ \varepsilon_{w_1, w_1 \cap w_2} = \{w_1 \cap w_2\}$, which then implies that $(w_1 \cap w_2) <_\varphi^{\text{exh}} w_1$. But, since w_1 and w_2 are models of μ and μ is a Horn formula, it follows that $(w_1 \cap w_2) \in [\mu]$. These facts, now, create a contradiction with the assumption that $w_1 \in \min_{\leq_\varphi^{\text{exh}}}[\mu]$.

("$\supseteq$") Take $w_1 \in \min_{\leq_\varphi^{\text{exh}*}}[\mu]$ and suppose that $w_1 \notin \min_{\leq_\varphi^{\text{exh}}}[\mu]$. This implies that there exists $w_2 \in [\mu]$ such that $w_2 <_\varphi^{\text{exh}} w_1$. But, since $\leq_\varphi^{\text{exh}*}$ is an ordering extension of $\leq_\varphi^{\text{exh}}$, this implies that $w_2 <_\varphi^{\text{exh}*} w_1$, which leads to a contradiction.

Proposition 6.3 make the same point as the one made by Theorem 6.3, but through the vehicle of Suzumura consistency, and provides another example of how rational choice can be of use to belief change.

Partial preorders

What happens if we replace postulate R_6 with the weaker postulates R_{7-8}? Experience teaches us that we should expect partial preorders represented via the exclusive revealed assignment. Switching to the Horn fragment complicates things, of course, but we can use Theorem 6.3 as a model for what HHH-revision for exclusive operators should look like. Example 6.7 shows that we need Horn compliance, on the semantic side, to make sure that an assignment on interpretations can represent an HHH-revision operator. And the case of exhaustive operators teaches us that another important detail is finding a revealed ranking on outcomes that can be extended to a transitive preorder, and that can be elicited using the postulates on hand.

In getting the relationship between the postulates and the ranking on outcomes right, the basic step is inferring the ranking on two interpretations. For exclusive operators, i.e., operators that satisfiy postulates R_{1-5} and R_{7-8}, the following lemma will prove crucial.

Lemma 6.1

If $\circ$ is an HHH-revision operator that satisfies postulates R_{1-5} and R_{7-8} then, for any interpretations w_1 and w_2 and an $\mathcal{L}_{\text{Horn}}$-proxy of $\{w_1, w_2\}$, it holds that $w_1 \in [\varphi \circ \varepsilon_{1,2}]$ if and only if $w_1 \in \min_{\leq_\varphi^{\text{exc}}} \text{Cl}_{\text{Horn}}(\{w_1, w_2\})$.

Proof

It is straightforward to see that the statement holds if $w_1 = w_2$. If $w_1 \neq w_2$ and $\{w_1, w_2\}$ is Horn-closed, i.e., $w_1 \subseteq w_2$ or $w_2 \subseteq w_1$, then $[\varepsilon_{1,2}] = \{w_1, w_2\}$, and the fact that $w_1 \in [\varphi \circ \varepsilon_{1,2}]$ is equivalent to the fact that either $w_1 <_\varphi^{\text{exc}} w_2$, in case $w_2 \notin [\varphi \circ \varepsilon_{1,2}]$, or that w_1 and w_2 are incomparable with respect to $\leq_\varphi^{\text{exc}}$, in case $w_2 \in [\varphi \circ \varepsilon_{1,2}]$. In either case, this is equivalent to w_1 being in $\min_{\leq_\varphi^{\text{exc}}} \text{Cl}_{\text{Horn}}(\{w_1, w_2\})$.

For the rest of the proof we will then assume that $w_1 \not\subseteq w_2$ and $w_2 \not\subseteq w_1$, which implies that $[\varepsilon_{1,2}] = \{w_0, w_1, w_2\}$, where $w_0 = w_1 \cap w_2$.

("$\Rightarrow$") Suppose $w_1 \in [\varphi \circ \varepsilon_{1,2}]$ and $w_1 \notin \min_{\leq_\varphi^{\text{exc}}} \text{Cl}_{\text{Horn}}(\{w_1, w_2\})$. The latter fact implies that there exists an interpretation in $\text{Cl}_{\text{Horn}}(\{w_1, w_2\})$ that is strictly better than w_1, i.e., $w_2 <_\varphi^{\text{exc}} w_1$ or $w_0 <_\varphi^{\text{exc}} w_1$. We show, by a case analysis, that this leads to a contradiction.

Case 1. If $w_2 <_\varphi^{\text{exc}} w_1$, then by the definition of the exclusive revealed assignment we have that $w_1 \notin [\varphi \circ \varepsilon_{1,2}]$, which is a contradiction.

Case 2 If $w_0 <_\varphi^{\text{exc}} w_1$, then, again, by the definition of the exclusive revealed assignment, we have that $w_1 \notin [\varphi \circ \varepsilon_{0,1}]$. However, we also have that:

$$\begin{aligned} (\varphi \circ \varepsilon_{1,2}) \wedge \varepsilon_{0,1} &\models \varphi \circ (\varepsilon_{1,2} \wedge \varepsilon_{0,1}) && \text{by } R_5 \\ &\equiv \varphi \circ \varepsilon_{0,1}. && \text{by } R_4 \end{aligned}$$

Since $w_1 \in [\varphi \circ \varepsilon_{1,2}]$ and $w_1 \in [\varepsilon_{0,1}]$, this implies that $w_1 \in [\varphi \circ \varepsilon_{0,1}]$ and we have thus arrived at a contradiction.

("$\Leftarrow$") Suppose $w_1 \in \min_{\leq_\varphi^{\text{exc}}} \text{Cl}_{\text{Horn}}(\{w_1, w_2\})$. and $w_1 \notin [\varphi \circ \varepsilon_{1,2}]$. Since, by postulates R_1 and R_3, it holds that $\emptyset \subset [\varphi \circ \varepsilon_{1,2}] \subseteq [\varepsilon_{1,2}]$, there are three possibilities for what $[\varepsilon_{1,2}]$ can be. We look at all the possibilities in turn.

Case 1. If $[\varphi \circ \varepsilon_{1,2}] = \{w_2\}$, then this implies, by the definition of the exclusive revealed assignmnent, that $w_2 <_\varphi^{\text{exc}} w_1$, which contradicts the fact that $w_1 \in \min_{\leq_\varphi^{\text{exc}}} \text{Cl}_{\text{Horn}}(\{w_1, w_2\})$.

Case 2. If $[\varphi \circ \varepsilon_{1,2}] = \{w_0, w_2\}$, then this also implies that $w_2 <_\varphi^{\text{exc}} w_1$, and the reasoning from Case 1 applies here as well.

Case 3. If $[\varphi \circ \varepsilon_{1,2}] = \{w_0\}$, then we have that:

$$\varphi \circ \varepsilon_{1,2} \models \varepsilon_{0,1},$$
$$\varphi \circ \varepsilon_{0,1} \models \varepsilon_{1,2},$$

which, by postulate R_6, implies that $\varphi \circ \varepsilon_{1,2} \equiv \varphi \circ \varepsilon_{0,1}$. In other words, it holds that $[\varphi \circ \varepsilon_{0,1}] = \{w_0\}$, which then implies that $w_0 <_\varphi^{\text{exc}} w_1$. But this contradicts the fact that $w_1 \in \min_{\leq_\varphi^{\text{exc}}} \text{Cl}_{\text{Horn}}(\{w_1, w_2\})$.

Though not immediately intuitive, Lemma 6.1 expresses a reassuring connection between the behavior of exclusive operators and the exclusive revealed assignment: it says that w_1 getting chosen when the choice set is $[\varepsilon_{1,2}]$ is equivalent to there not being an interpretation in $[\varepsilon_{1,2}]$ that is strictly better than w_1 according to $\leq_\varphi^{\text{exc}}$. While in normal circumstances such a statement would be quasi-obvious, in this case we could become convinced of it only with some effort. As such, Lemma 6.1 validates the usefulness both of $\mathcal{L}_{\text{Horn}}$-proxies of $\{w_1, w_2\}$ and of the exclusive revealed assignment for comparing w_1 and w_2. In other words, it shows that the two notions can work together with exclusive HHH-revision operators in emulating rational choice behavior—at least when it comes to two interpretations.

The same result, however, hints to the kind of property that has to be in place for longer chains of comparisons. For exhaustive operators we used Suzumura consistency to make sure that the revealed assignments avoids unwanted structure. For exclusive operators, the following property turns out to be required. It is expected to hold for any integer $n \geq 1$ and pairwise distinct interpretations $w_1, \ldots, w_n$:

(r_{HSC}) If $w_1 <_\varphi \cdots <_\varphi w_n$, then $w_n \notin \min_{\leq_\varphi} \text{Cl}_{\text{Horn}}(\{w_1, w_n\})$.

Property r_{HSC}, where 'HSC' stands for *Horn Suzumura consistency*, is an adaptation of Suzumura consistency to the context of partial preorders and the Horn fragment, and is intended to apply to the exclusive revealed assignment, inferred using an exclusive

revision operator. Through Lemma 6.1, it is readily apparent that property $\mathrm{r}_{\mathrm{HSC}}$ still embodies the spirit of traditional Suzumura consistency: in plain words, it says that if there is a chain of comparisons that starts with w_1 and ends with w_n, then w_n should not be chosen by the revision function when given a choice between w_1 and w_n. The expression of the property is complicated, in this case, by the fact that the choice between w_1 and w_n, in the Horn fragment, may be done through a choice set that includes $w_1 \cap w_n$ as well.

Property $\mathrm{r}_{\mathrm{HSC}}$ must be complemented on the syntactic side by a postulate, here a postulate schema, which is expected to hold for any Horn formula φ, intepretations w_1, $\ldots$, w_n and the corresponding $\mathcal{L}_{\mathrm{Horn}}$ proxies:

($\mathrm{R}_{\mathrm{HSC}}$) If $(\varphi \circ \varepsilon_{1,2}) \wedge \varepsilon_1$ is consistent and $(\varphi \circ \varepsilon_{1,2}) \wedge \varepsilon_2$ is inconsistent, $\ldots$, $(\varphi \circ \varepsilon_{n-1,n}) \wedge \varepsilon_{n-1}$ is consistent and $(\varphi \circ \varepsilon_{n-1,n}) \wedge \varepsilon_n$ is inconsistent, then $(\varphi \circ \varepsilon_{n,1}) \wedge \varepsilon_n$ is inconsistent.

Postulate $\mathrm{R}_{\mathrm{HSC}}$ expresses the same idea as property $\mathrm{r}_{\mathrm{HSC}}$, but using the formulas and the revision operator as a choice device, instead of the preorder. It can be readily seen that, at least insofar as the exclusive revealed assignment is concerned, postulate $\mathrm{R}_{\mathrm{HSC}}$ and property $\mathrm{r}_{\mathrm{HSC}}$ go hand in hand.

Theorem 6.4

If $\circ$ is an HHH-revision operator that satisfies postulates R_{1-5} and R_{7-8}, then $\circ$ satisfies postulate $\mathrm{R}_{\mathrm{HSC}}$ if and only if the exclusive $\circ$-revealed $\mathcal{L}_{\mathrm{Horn}}$-assignment $\preccurlyeq^{\mathrm{exc}}$ satisfies property $\mathrm{r}_{\mathrm{HSC}}$.

Proof

("$\Rightarrow$") Take an HHH-revision operator $\circ$ that satisfies postulates R_{1-5}, R_{7-8} and $\mathrm{R}_{\mathrm{HSC}}$, and suppose there exists a formula φ and pairwise distinct interpretations w_1, $\ldots$, w_n such that $w_1 <_\varphi^{\mathrm{exc}} \cdots <_\varphi^{\mathrm{exc}} w_n$. By the definition of $\leq_\varphi^{\mathrm{exc}}$, this implies that $(\varphi \circ \varepsilon_{1,2}) \wedge \varepsilon_1$ is consistent, $\ldots$, $(\varphi \circ \varepsilon_{n-1,n}) \wedge \varepsilon_{n-1}$ is consistent. Suppose, in addition, that $w_n \in \min_{\leq_\varphi^{\mathrm{exc}}} \mathrm{Cl}_{\mathrm{Horn}}(\{w_1, w_n\})$. By Lemma 6.1, it follows from this that $w_n \in [\varphi \circ \varepsilon_{1,n}]$, i.e., that $(\varphi \circ \varepsilon_{1,n}) \wedge \varepsilon_n$ is consistent. But this contradicts the fact that $\circ$ satisfies postulate $\mathrm{R}_{\mathrm{HSC}}$.

("$\Leftarrow$") Suppose $(\varphi \circ \varepsilon_{1,2}) \wedge \varepsilon_1$ is consistent, $\ldots$, $(\varphi \circ \varepsilon_{n-1,n}) \wedge \varepsilon_{n-1}$ is consistent. This implies that $w_1 <_\varphi^{\mathrm{exc}} \cdots <_\varphi^{\mathrm{exc}} w_n$. Since $\leq_\varphi^{\mathrm{exc}}$ satisfies property $\mathrm{r}_{\mathrm{HSC}}$, it follows that $w_n \notin \min_{\leq_\varphi^{\mathrm{exc}}} \mathrm{Cl}_{\mathrm{Horn}}\{w_1, w_n\}$, which means, using Lemma 6.1, that $(\varphi \circ \varepsilon_{1,n}) \wedge \varepsilon_n$ is inconsistent.

Postulate $\mathrm{R}_{\mathrm{HSC}}$ is similar to postulate R_{SC}, used for exhaustive HHH-revision operators, so one might wonder whether R_{SC} could not be used here as well. The following example,

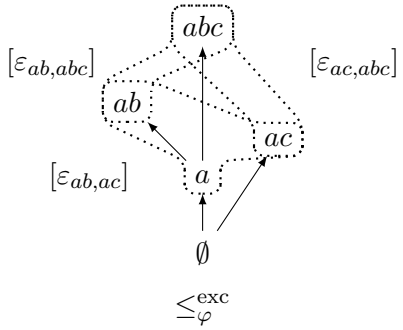


Figure 6.5: Exclusive $\circ$-revealed preorder $\leq_\varphi^{\text{exc}}$ that is invalidated by postulate R_{SC}, even though we would like to see $\leq_\varphi^{\text{exc}}$ allowed.

however, shows that when we are working with partial preorders, postulate R_{SC} is not quite the right choice.

Example 6.9: Postulate R_{SC} does not work for partial preorders

Consider an HHH-revision operator $\circ$ and a Horn formula φ with $[\varphi] = \{\emptyset\}$, for which $[\varphi \circ \varepsilon_{ab,abc}] = \{ab, abc\}$, $[\varphi \circ \varepsilon_{abc,ac}] = \{abc, ac\}$ $[\varphi \circ \varepsilon_{ab,ac}] = \{a, ac\}$ and $\varphi \circ \varepsilon_{\emptyset,w} = \{\emptyset\}$, for any interpretation w. Note that this set of revision results is perfectly consistent with postulates R_5 and R_{7-8} and, furthermore, induces the following exclusive ranking $\leq_\varphi^{\text{exc}}$: ab, abc and ac are pairwise $\leq_\varphi^{\text{exc}}$-incomparable, with the same going for a and ac, $a <_\varphi^{\text{exc}} ab$. The preorder $\leq_\varphi^{\text{exc}}$ is depicted in Figure 6.5. This is a legitimate partial preorder, that should be allowed. However, under postulate R_{SC}, this setup is disallowed. Note, we have that:

$$
\begin{aligned}
[(\varphi \circ \varepsilon_{ab,abc}) \wedge \varepsilon_{ab}] &= \{ab\},\\
[(\varphi \circ \varepsilon_{abc,ac}) \wedge \varepsilon_{abc}] &= \{abc\},\\
[(\varphi \circ \varepsilon_{ac,ab}) \wedge \varepsilon_{ac}] &= \{ac\},\\
[(\varphi \circ \varepsilon_{ac,ab}) \wedge \varepsilon_{ab}] &= \emptyset.
\end{aligned}
$$

Thus, if we take $w_1 = ab$, $w_2 = abc$ and $w_3 = ac$, the setup summarized in Figure 6.5 would constitute a counter-example to postulate R_{SC}, since in this case postulate R_{SC} implies that it is not possible to have $(\varphi \circ \varepsilon_{ac,ab}) \wedge \varepsilon_{ac}$ consistent and $(\varphi \circ \varepsilon_{ac,ab}) \wedge \varepsilon_{ab}$ inconsistent.

Example 6.9 shows that postulate R_{SC}, in conjunction with postulates R_{1-5} and R_{7-8}, leads to unwanted effects, as it eliminates preorders that we would like to allow. Postulate R_{HSC} is the postulate we need for this case.

That being said, property r_{HSC}, which postulate R_{HSC} guarantees, is still closely connected to Suzumura consistency, which property r_{SC} encodes. In particular, r_{HSC} can be seen as a more specialized version of Suzumura consistency, tweaked to work with preorders

revealed by HHH-revision operators. Proposition 6.4 shows that, in this context, property r_{HSC} actually implies Suzumura consistency.

Proposition 6.4

If $\circ$ is an HHH-revision operator that satisfies postulates R_{1-5} and R_{7-8}, and the exclusive revealed assignment $\preccurlyeq^{exc}$ satisfies property r_{HSC}, then $\preccurlyeq^{exc}$ also satisfies property r_{SC}.

Proof

Recall that, in the context of an exclusive revealed assignment $\preccurlyeq^{exc}$, a preference order $\leq^{exc}_{\varphi}$ is strict for any distinct interpretations w_1 and w_2. Take, then, pairwise distinct interpretations $w_1, \ldots, w_n$ such that $w_1 <^{exc}_{\varphi} \cdots <^{exc}_{\varphi} w_n$ and suppose, in addition, that $w_n <^{exc}_{\varphi} w_1$. By the definition of the exclusive revealed assignment, this means that $w_n \in [\varphi \circ \varepsilon_{1,n}]$, which, by Lemma 6.1, implies that $w_n \in \min_{\leq^{exc}_{\varphi}} Cl_{Horn}\{w_1, w_n\}$. But this contradicts the fact that $\leq^{exc}_{\varphi}$ satisfies property r_{HSC}.

One of the effects of Proposition 6.4, when combined with Theorem 2.1, implies that the exclusive revealed preorders of an exclusive HHH-revision operator can also be extended to total preorders on outcomes, as long as the assignment can be shown to satisfy property r_{HSC}. For our purposes, though, total preorders are overkill, since partial preorders are all we are looking for. Theorem 6.5 shows that, given all the conceptual legwork done so far, such partial preorders is within easy reach: though not exactly the exclusive revealed rankings, they can be elicited from them by the transitive closure.

Theorem 6.5

An HHH-revision operator $\circ$ satisfies postulates R_{1-5}, R_{7-8} (i.e., is exclusive) and postulate R_{HSC} if and only if there exists an $\mathcal{L}_{Horn}$-assignment $\preccurlyeq$ on interpretations that satisfies properties r_{1-2}, r_4, r_{6-7} and r_{HC} (i.e., is partial, syntax independent, r-faithful and Horn compliant) and that represents the operator $\circ$.

Proof

("$\Leftarrow$") Take, first, an assignment $\preccurlyeq$ that satisfies the specified conditions. Since $\preccurlyeq$ is Horn compliant, this implies that the $\preccurlyeq$-induced operator $\circ^{\preccurlyeq}$ is an HHH-revision operator. Checking that $\circ^{\preccurlyeq}$ satisfies postulates R_{1-5}, R_{7-8} is routine. Satisfaction of postulate R_{HSC} follows using Theorem 6.4.

("$\Rightarrow$") Take, now, an HHH-revision operator $\circ$ that satisfies postulates R_{1-5}, R_{7-8} and R_{HSC}, and the $\circ$-revealed exclusive $\mathcal{L}_{Horn}$-assignment $\preccurlyeq^{exc}$. If φ is a Horn formula, then the preorder $\leq^{exc}_{\varphi}$ is neither complete, nor transitive. We know, however, by

Theorem 6.4, that $\leq_\varphi^{\text{exc}}$ satisfies property r_{HSC}.

We take $\leq_\varphi^*$ to be the transitive and reflexive closure of $\leq_\varphi^{\text{exc}}$, and the assignment $\preccurlyeq^*$ to be defined by $\preccurlyeq^*(\varphi) \stackrel{\text{def}}{=} \leq_\varphi^*$. We show, now, that $\preccurlyeq^*$ is the assignment we are looking for. We have that $\preccurlyeq^*$ is reflexive and transitive by definition, and using postulate R_2 it can be shown in the usual way that $\preccurlyeq^*$ is r-faithful. The only thing left to be shown is that $\preccurlyeq^*$ represents $\circ$, i.e., that $[\varphi \circ \mu] = \min_{\leq_\varphi^*}[\mu]$. We show this by double inclusion.

("$\subseteq$") Take $w \in [\varphi \circ \mu]$ and suppose $w \notin \min_{\leq_\varphi^*}[\mu]$. This means that there exists $w' \in [\mu]$ such that $w' <_\varphi^* w$, i.e., that there exist pairwise distinct interpretations w_0, $\ldots$, w_k such that:

$$w' <_\varphi^{\text{exc}} w_0 <_\varphi^{\text{exc}} \cdots <_\varphi^{\text{exc}} w_k <_\varphi^{\text{exc}} w.$$

Since $\leq_\varphi^{\text{exc}}$ satisfies property r_{HSC}, it follows that $w \notin \min_{\leq_\varphi^{\text{exc}}}\text{Cl}_{\text{Horn}}(\{w_1, w_n\})$. At the same time, we have that:

$$\begin{aligned}(\varphi \circ \mu) \wedge \varepsilon_{w,w'} &\models \varphi \circ (\mu \wedge \varepsilon_{w,w'}) && \text{by } \text{R}_5\\ &\equiv \varphi \circ \varepsilon_{w,w'}. && \text{by } \text{R}_4\end{aligned}$$

Since $w \in [\varphi \circ \mu]$, this implies that $w \in [\varphi \circ \varepsilon_{w,w'}]$ and, using Lemma 6.1 we conclude that $w \in \min_{\leq_\varphi^{\text{exc}}}\text{Cl}_{\text{Horn}}(\{w_1, w_n\})$, which creates a contradiction.

("$\supseteq$") Take $w \in \min_{\leq_\varphi^*}[\mu]$ and an arbitrary interpretation $w' \in [\mu]$. We will show first that $w \in [\varphi \circ \varepsilon_{w,w'}]$. Suppose that $w \notin [\varphi \circ \varepsilon_{w,w'}]$. Since $[\varepsilon_{w,w'}] = \text{Cl}_{\text{Horn}}(\{w, w'\})$, there are two relevant cases to look at.

Case 1. If $[\varphi \circ \varepsilon_{w,w'}] = \{w'\}$ or $[\varphi \circ \varepsilon_{w,w'}] = \{w', w \cap w'\}$, then, by the definition of $\leq_\varphi^{\text{exc}}$ it follows that $w' <_\varphi^{\text{exc}} w$, which implies that $w' <_\varphi^* w$, contradicting the fact that $w \in \min_{\leq_\varphi^*}[\mu]$.

Case 2. If $[\varphi \circ \varepsilon_{w,w'}] = \{w \cap w'\}$, then we have that:

$$\begin{aligned}\varphi \circ \varepsilon_{w,w\cap w'} &\models \varepsilon_{w,w'},\\ \varphi \circ \varepsilon_{w,w'} &\models \varepsilon_{w,w\cap w'}.\end{aligned}$$

Using postulate R_7, it follows that $\varphi \circ \varepsilon_{w,w\cap w'} \equiv \varphi \circ \varepsilon_{w,w'}$, i.e., $[\varphi \circ \varepsilon_{w,w\cap w'}] = \{w \cap w'\}$. This, now, implies that $(w \cap w') <_\varphi^{\text{exc}} w$, from which it follows that $(w \cap w') <_\varphi^* w$. Since $w \in [\mu]$, $w' \in [\mu]$ and the fact that μ is a Horn formula and thus $[\mu]$ is closed under intersection, we infer that $(w \cap w') \in [\mu]$ as well. But this, together with the previously inferred statement that $(w \cap w') <_\varphi^* w$, contradicts the fact that $w \in \min_{\leq_\varphi^*}[\mu]$. We conclude, therefore, that $w \in [\varphi \circ \varepsilon_{w,w'}]$.

For the final step, we use postulate R_8 and the fact that $\mu = \bigvee_{w' \in [\mu]} \varepsilon_{w,w'}$ to conclude that $w \in [\varphi \circ (\bigvee_{w' \in [\mu]} \varepsilon_{w,w'})]$.

Theorem 6.5 shows that partial preorders can be used to represent exclusive HHH-revision operators, just like total preorders can be used to represent exhaustive HHH-revision operators. In doing so, some additions need to be made, both on the semantic side and on the syntactic side. Firstly, Horn compliance works with both total and partial preorders, and guarantees that revision always falls within the Horn fragment. Then, Suzumura consistency, together with its accompanying postulate, kicks in to make sure that the revealed assignments do not contain cycles or other unwanted side effects. Interestingly, Suzumura consistency needs to be slightly modified in order to work for partial preorders and postulates R_{1-5} and R_{7-8}.

6.4 Horn update by Horn formulas

All the wisdom gained in Section 6.3 can be put to use to understand update with Horn formulas, and this section is dedicated to spelling out the details. Since the primary plot points for Horn update are the same as for Horn revision with Horn formulas, we defer to the previous section for discussion and motivation.

An HHH-*update operator* $\diamond$ is a function $\diamond\colon \mathcal{L}_{\mathrm{Horn}} \times \mathcal{L}_{\mathrm{Horn}} \to \mathcal{L}_{\mathrm{Horn}}$, taking as input two Horn formulas, typically denoted by φ and μ and referred to as the prior and new information, respectively, and returning a Horn formula, typically denoted by $\varphi \diamond \mu$ and referred to as the posterior updated information. The postulates for HHH-update revision are the standard update postulates U_{1-9} as presented in Section 3.2, and particularized, as for revision, to Horn formulas. The postulates are expected to apply to any Horn formulas φ, φ_1, φ_2, μ, μ_1 and μ_2 and complete Horn formulas $\dot{\varphi}$:

(U_1) $\varphi \diamond \mu \models \mu$.

(U_2) If $\varphi \models \mu$, then $\varphi \diamond \mu \equiv \varphi$.

(U_3) If φ and μ are consistent, then $\varphi \diamond \mu$ is consistent.

(U_4) If $\varphi_1 \equiv \varphi_2$ and $\mu_1 \equiv \mu_2$, then $\varphi_1 \diamond \mu_1 \equiv \varphi_2 \diamond \mu_2$.

(U_5) $(\varphi \diamond \mu_1) \wedge \mu_2 \models \varphi \diamond (\mu_1 \wedge \mu_2)$.

(U_6) If $(\dot{\varphi} \diamond \mu_1) \wedge \mu_2$ is consistent, then $\dot{\varphi} \diamond (\mu_1 \wedge \mu_2) \models (\dot{\varphi} \diamond \mu_1) \wedge \mu_2$.

(U_7) If $\varphi \diamond \mu_1 \models \mu_2$ and $\varphi \diamond \mu_2 \models \mu_1$, then $\varphi \diamond \mu_1 \equiv \varphi \diamond \mu_2$

(U_8) If $\mu \equiv \mu_1 \vee \mu_2$. then $(\dot{\varphi} \diamond \mu_1) \wedge (\dot{\varphi} \diamond \mu_2) \models \dot{\varphi} \diamond \mu$.

(U_9) If $\varphi \equiv \varphi_1 \vee \varphi_2$, then $(\varphi_1 \vee \varphi_2) \diamond \mu \equiv (\varphi_1 \diamond \mu) \vee (\varphi_2 \diamond \mu)$.

As for $\mathcal{L}$-update, we are more interested in the variant of postulate U_9 presented below. Since we are working in the Horn fragment, this means that proxies are intended to be Horn formulas: in other words, if v is an interpretations, then ε_v is an $\mathcal{L}_{\mathrm{Horn}}$-proxy of $\{v\}$, i.e., a Horn formula such that $[\varepsilon_v] = \mathrm{Cl}_{\mathrm{Horn}}(\{v\}) = \{v\}$.

(U$_{10}$) $\varphi \diamond \mu \equiv \bigvee_{v \in [\varphi]} (\varepsilon_v \diamond \mu)$.

Postulate U$_{10}$ is equivalent to postulate U$_9$ even when applied to Horn formulas, and says that $\varphi \diamond \mu$ can be decomposed in the results for $\varepsilon_v \diamond \mu$, for every $v \in [\varphi]$.

On the semantic side we can use, as for $\mathcal{L}$-update, assignments on complete formulas: every complete formula is, or can be thought of, as a Horn formula, in the sense that its set of models is Horn-closed. Thus, we can work here with n $\mathcal{L}_{\text{comp}}$ *assignment* $\preccurlyeq$ *on interpretations*, which is a function $\preccurlyeq\colon \mathcal{L}_{\text{comp}} \to 2^{\mathcal{U} \times \mathcal{U}}$, taking as input a complete formula $\dot{\varphi}$ and returning a binary relation on interpretations. The properties we are interested in are as follows, for any complete Horn propositional formulas $\dot{\varphi}$, $\dot{\varphi}_1$, $\dot{\varphi}_2$ and interpretations w, v, w_1 and w_2:

(u$_1$) $w \leq_{\dot{\varphi}} w$.

(u$_2$) If $w_1 \leq_{\dot{\varphi}} w_2$ and $w_2 \leq_{\dot{\varphi}} w_3$, then $w_1 \leq_{\dot{\varphi}} w_3$.

(u$_3$) $w_1 \leq_{\dot{\varphi}} w_2$ or $w_2 \leq_{\dot{\varphi}} w_1$.

(u$_4$) If $\dot{\varphi}_1 \equiv \dot{\varphi}_2$, then $\leq_{\dot{\varphi}_1} = \leq_{\dot{\varphi}_2}$.

(u$_5$) If $[\dot{\varphi}] = \{v\}$ and $w \neq v$, then $v <_{\dot{\varphi}} w$.

As usual, an $\mathcal{L}_{\text{comp}}$-assignment $\preccurlyeq$ on interpretations is *partial* if it satisfies properties u$_{1-2}$, *total* if it satisfies properties u$_{1-3}$, *syntax insensitive* if it satisfies property u$_4$ and *u-faithful* if it satisfies property u$_5$.

If $\diamond$ is an HHH-update operator and $\preccurlyeq$ is an $\mathcal{L}_{\text{comp}}$-assignment on interpretations, then $\preccurlyeq$ *represents* $\diamond$ (and $\diamond$ *is represented by* $\preccurlyeq$) if, for any Horn formulas φ and μ, it holds that $[\varphi \diamond \mu] = \bigcup_{v \in [\varphi]} \min_{\leq_{\varepsilon_v}} [\mu]$. Given an $\mathcal{L}_{\text{comp}}$-assignment $\preccurlyeq$ on interpretations, the $\preccurlyeq$*-induced* $\mathcal{L}$*-update operator* $\diamond^{\preccurlyeq}$ is defined, for any Horn formulas φ and μ, by taking:

$$[\varphi \diamond^{\preccurlyeq} \mu] \stackrel{\text{def}}{=} \bigcup_{v \in [\varphi]} \min_{\leq_{\varepsilon_v}} [\mu].$$

If $\diamond$ is an HHH-update operator and $\dot{\varphi}$ is a complete (Horn) formula, the *exhaustive* $\diamond$*-revealed plausibility relation* $\leq^{\text{exh}}_{\dot{\varphi}}$ and the *exclusive* $\diamond$*-revealed plausibility relation* $\leq^{\text{exc}}_{\dot{\varphi}}$ are defined, for any interpretations w_1 and w_2, respectively, as:

$$w_1 \leq^{\text{exh}}_{\dot{\varphi}} w_2 \text{ if } w_1 \in [\dot{\varphi} \diamond \varepsilon_{1,2}],$$
$$w_1 \leq^{\text{exc}}_{\dot{\varphi}} w_2 \text{ if either } w_1 = w_2, \text{or } w_1 \in [\dot{\varphi} \diamond \varepsilon_{1,2}] \text{ and } w_2 \notin [\dot{\varphi} \diamond \varepsilon_{1,2}].$$

The *exhaustive* $\diamond$*-revealed* $\mathcal{L}_{\text{comp}}$*-assignment* $\preccurlyeq^{\text{exh}}$ and *exclusive* $\diamond$*-revealed* $\mathcal{L}_{\text{comp}}$*-assignment* $\preccurlyeq^{\text{exc}}$ are obtained by taking $\preccurlyeq^{\text{exh}}(\dot{\varphi}) = \leq^{\text{exh}}_{\dot{\varphi}}$ and $\preccurlyeq^{\text{exc}}(\dot{\varphi}) = \leq^{\text{exc}}_{\dot{\varphi}}$, for any complete Horn formula $\dot{\varphi}$. As for revision, $\diamond^{\preccurlyeq}$ is defined as an $\mathcal{L}$-update operator, i.e., an operator that

returns propositional formulas and not necessarily Horn formula, because an assignment needs to be restricted if it is to deliver results compatible with the Horn fragment.

Since update becomes indistinguishable from revision when φ is complete, the same problems that plague HHH-revision also occur when working with HHH-update operators. Note, for instance, that the prior information in Examples 6.7 is complete: this shows that this example is relevant to HHH-update as well and, in particular, that existing update operators, such as Forbus and Winslett's operators do not work as HHH-update operators. The solution, of course, is to restrict the assignments we are working with using a property that, as we might expect, is a version of Horn compliance adapted to the context of update. This property says that the following must hold, for any Horn formulas φ and μ:

($\mathrm{u_{HC}}$) $\bigcup_{v\in[\varphi]} \min_{\leq_{\varepsilon_v}}[\mu]$ is Horn-closed.

Property $\mathrm{u_{HC}}$, where 'HC' stands for *Horn compliance*, works along the same lines as Horn compliance in the context of HHH-revision [Delgrande and Peppas, 2015], and ensures that the $\preccurlyeq$-induced update operator is an HHH-update operator.

Furthermore, Examples 6.8 and 6.9 can also be adapted to HHH-update operator, showing that the standard postulates do not exclude unwanted assignments. The fixes, as for HHH-revision, involve a series of semantic properties on the revealed preference orders, coupled with corresponding logical postulates. The semantic properties are intended to apply for any complete Horn formulas $\dot{\varphi}$ and interpretations $w_1, \ldots, w_n$, and are as follows:

($\mathrm{u_{SC}}$) If $w_1 \leq_{\dot{\varphi}} \cdots \leq_{\dot{\varphi}} w_n$, then $w_n \not<_{\dot{\varphi}} w_1$.

($\mathrm{u_{HSC}}$) If $w_1 <_{\dot{\varphi}} \cdots <_{\dot{\varphi}} w_n$, then $w_n \notin \min_{\leq_{\dot{\varphi}}} \mathrm{Cl_{Horn}}(\{w_1, w_n\})$.

These properties are equivalent to properties $\mathrm{r_{SC}}$ and $\mathrm{r_{HSC}}$, with the only difference being that they are particularized to complete Horn formulas. As expected, property $\mathrm{u_{SC}}$ is expected to hold for the exhaustive revealed assignment and $\mathrm{u_{HSC}}$ is intended to hold for the exclusive revealed assignment.

On the syntactic side we use the following postulates, or more precesily postulate schemas, intended to work for any complete Horn formulas $\dot{\varphi}$, interpretations $w_1, \ldots, w_n$ and the corresponding $\mathcal{L}_{\mathrm{Horn}}$-proxies:

($\mathrm{U_{SC}}$) If $(\dot{\varphi} \diamond \varepsilon_{1,2}) \wedge \varepsilon_1$ is consistent, $\ldots$, $(\dot{\varphi} \diamond \varepsilon_{n-1,n}) \wedge \varepsilon_{n-1}$ is consistent, then it does not hold both that $(\dot{\varphi} \diamond \varepsilon_{n,1}) \wedge \varepsilon_n$ is consistent and that $(\dot{\varphi} \diamond \varepsilon_{n,1}) \wedge \varepsilon_1$ is inconsistent.

($\mathrm{U_{HSC}}$) If $(\dot{\varphi} \diamond \varepsilon_{1,2}) \wedge \varepsilon_1$ is consistent and $(\dot{\varphi} \diamond \varepsilon_{1,2}) \wedge \varepsilon_2$ is inconsistent, $\ldots$, $(\dot{\varphi} \diamond \varepsilon_{n-1,n}) \wedge \varepsilon_{n-1}$ is consistent and $(\dot{\varphi} \diamond \varepsilon_{n-1,n}) \wedge \varepsilon_n$ is inconsistent, then $(\dot{\varphi} \diamond \varepsilon_{n,1}) \wedge \varepsilon_n$ is inconsistent.

Postulates U_{SC} and U_{HSC} are direct rewritings of postulates R_{SC} and R_{HSC} from HHH-revision, and the argument that they go hand in hand with propeerties u_{SC} and u_{HSC} is entirely similar.

With most of the theoretical groundwork having been laid in Section 6.3, the representation results can be stated now in quick succession. The first result concerns HHH-update operators that satisfy postulates U_{1-6}, U_9 and U_{SC}, and are represented using total assignments.

Theorem 6.6

An HHH-update operator $\diamond$ satisfies postulates U_{1-6}, U_9 and U_{SC} if and only if there exists an $\mathcal{L}_{comp}$-assignment $\preccurlyeq$ on interpretations that satisfies properties u_{1-5} and u_{HC} (i.e., is total, syntax independent, u-faithful and Horn compliant) and that represents the operator $\diamond$.

Proof

Horn compliance of $\preccurlyeq$ guarantees that the $\preccurlyeq$-induced update operator $\diamond^{\preccurlyeq}$ is an HHH-update operator, and showing that $\diamond^{\preccurlyeq}$ satisfies postulates U_{1-6}, U_9 and U_{SC} is straightforward, and works in the same way as for HHH-revision operators. Conversely, if $\diamond$ satisfies postulates U_{1-6}, U_9 and U_{SC}, then we use the argument for HHH-revision operators that satisfy postulates R_{1-6}, particularized to complete formulas. That is to say, using Theorem 6.3, we know that the $\diamond$-revealed exhaustive assignment represents $\diamond$ for the case of complete formulas, i.e., if $\dot{\varphi}$ is a complete Horn formula, then $[\dot{\varphi} \diamond \mu] = \min_{\leq_{\dot{\varphi}}}[\mu]$, for any Horn formula μ. We now use postulate U_9 to extend this to the fact that $[\varphi \diamond \mu] = \bigcup_{v \in [\varphi]} \min_{\leq_{\varepsilon_v}}[\mu]$, for any Horn formula φ. This shows that $\preccurlyeq^{exh}$ represents the operator $\diamond$.

Accompanying this result is, of course, a result for HHH-update operators that satisfy postulates U_{1-5}, U_{7-8}, U_9 and U_{HSC}, and are represented using partial assignments.

Theorem 6.7

An HHH-update operator $\diamond$ satisfies postulates U_{1-5}, U_{7-8}, U_9 and U_{HSC} if and only if there exists an $\mathcal{L}_{comp}$-assignment $\preccurlyeq$ on interpretations that satisfies properties u_{1-2}, u_{4-5} and u_{HC} (i.e., is partial, syntax independent, u-faithful and Horn compliant) and that represents the operator $\diamond$.

Proof

The proof here follows the same lines as the proof for Theorem 6.6, but using the results in Theorem 6.5 to construct the revealed assignment that ends up representing

the operator $\diamond$.

Theorems 6.6 and 6.7 are straightforward applications of Theorems 6.3 and 6.5, but the insights gained in making these results work show that they are robust enough to function across different contexts.

6.5 Related work

Work on belief change for the Horn fragment began with work on contraction [Delgrande, 2008, Delgrande and Wassermann, 2010, Delgrande and Wassermann, 2013], the conclusions of which, however, were that traditional AGM techniques, such as ones based on remainder sets, did not work as expected in the Horn fragment. Alternative constructions were proposed, e.g., in terms of *weak remainder sets*, which were later taken up and extended [Booth et al., 2009, Booth et al., 2011, Zhuang and Pagnucco, 2014].

Getting Horn contraction right is important for revision as well, since contraction and revision are traditionally thought to be inter-definable via the Levi and Harper identities [Levi, 1991, Harper, 1976]. However, these identities involve negations and, since the negation of a Horn formula is not necessarily a Horn formula, their application to the Horn fragment is limited [Delgrande, 2008, Zhuang et al., 2013, Zhuang et al., 2017].

Thus, it seems that Horn revision is best thought of on its own. In this, some of the most promising approaches to revision of Horn formulas have been model based: given our understanding of revision as a choice function on outcomes, this line of research is particularly apposite. Indeed, our starting point has been the model developed originally by James Delgrande and Pavlos Peppas [Delgrande and Peppas, 2015, Delgrande et al., 2018], which we have sought to emulate here using partial preorders and weaker versions of the classical postulates. Our understanding of this model and its variations is thoroughly semantic, with the main actors, and guarantors of rationality, being the preorders on interpretations: in our view, as long as we are able to get the right kinds of preorders (i.e., transitive, or at least non-cyclic), then we are on the right track. In this light, the role of the additional postulates is to make sure that the preorders behave well: it is difficult to justify a postulate such as $\mathrm{R_{SC}}$ or $\mathrm{R_{HSC}}$ other than by appeal to the semantic property it induces. This is also why we choose to formulate these postulates using $\mathcal{L}_{\mathrm{Horn}}$-proxies (i.e., as choices over pairs of interpretations) rather than by using generic Horn formulas, as in [Delgrande and Peppas, 2015, Delgrande et al., 2018].

This preorder-led view is also what led us to the rational choice literature, where the relationship between choice functions and various types of preference relations has been a mainstay since the very early days. Suzumura consistency [Suzumura, 1976, Suzumura, 1983, Bossert and Suzumura, 2010] jumped out as the most obvious connecting element, but the rational choice literature is awash in distinctions and properties that could be useful to belief change as well. For instance, the acyclicity postulate presented in [Delgrande and Peppas, 2015, Delgrande et al., 2018] occurs in an equivalent formulation

in rational choice theory, as the *Strong Axiom of Revealed Preference (SARP)* [Hansson, 1968, Suzumura, 2016].

Belief change in the Horn fragment, as we have seen, is caught in between two equally demanding requirements: one is the normative demands provided by the postulates, i.e., the requirement that the revision operator should satisfy a set of desirable properties; the other is the expressibility requirement, i.e., we need to make sure that the result can be expressed as a Horn formula. In this chapter we have typically opted to hold fast to the postulates, and even complement them if needed, but this has had the effect of limiting the range of available operators: none of the tried and tested revision or update operators works in the Horn fragment. A different approach is to take the existing operators and repair their output when it does not fit into the Horn fragment, e.g., by taking the Horn-closure of the selected set of interpretations as the result. The advantage of this method is that it is guaranteed to produce Horn-closed results; the disadvantage is that some of the postulates might not be satisfied [Creignou et al., 2014, Creignou et al., 2016, Creignou et al., 2018b].

It should also be mentioned that belief change in fragments is not confined to contraction and revision (or update), with existing work on merging in the Horn fragment [Haret et al., 2015, Haret et al., 2017], and the Horn fragment is not the only formalism of interest, with logic programs, Answer Set Programs and Description Logics all enjoying their own fifteen minutes of AGM fame [Delgrande and Wassermann, 2013, Zhuang et al., 2016, Binnewies et al., 2018, Zhuang et al., 2019, Zheleznyakov et al., 2020].

6.6 Conclusion

In this chapter we have looked at revision of Horn formulas using both propositional and Horn formulas: though the results have ultimately taken a familiar form (i.e., as representation theorems), the overall work surrounding them shows that restricting the language in which belief change occurs does not allow us to seamlessly derive the same conclusions.

Over and over, we have seen that the standard postulates do not generalize immediately to the Horn fragment. For revision with propositional formulas, the standard postulate R_2 becomes problematic and the newly introduced neutrality postulate $\mathrm{R}_{\mathsf{NEUT}}$ proves impossible to satisfy. For revision, as well as update, with Horn formulas, it is postulates R_{5-6}, and, respectively, U_{5-6}, that end up being less powerful at forcing the assignments into the right shape. The solution, we have seen, has been to complement the standard postulates by new ones: typically, a special postulate depending on whether we were working with total or partial preorders. In doing so, we extended existing work on revision in the Horn fragment [Delgrande and Peppas, 2015, Delgrande et al., 2018], and revealed some interesting principles that underlie belief change in fragments.

Lemma 6.1, in particular, is interesting enough that it deserves more consideration: in essence, it says that if a chain of comparisons favors an interpretation w_1 over w_n, then

w_n should not be chosen over w_1 when the choice is over the proxy of w_1 and w_n. This property holds for all the settings we have encountered so far (total and partial orders, propositional logic as well as its Horn fragment), but it was only for the Horn fragment with partial preorders that we had to spell it out explicitly. As such, Lemma 6.1 captures a property that is likely to prove important for belief change in formalisms other than propositional logic and the Horn fragment.

CHAPTER 7

Preference Change

Preferences play a central role in theories of decision making as part of the mechanism underlying intentional behavior and rational choice: they show up in economic models of rational agency [Mas-Colell et al., 1995, Sen, 2017], as well as in formal models of artificial agents supposed to interact with the world and each other [Boutilier et al., 2004, Domshlak et al., 2011, Rossi et al., 2011, Pigozzi et al., 2016]. Since such interactions take place in dynamic environments, it can be expected that preferences change in response to new developments.

In this chapter we are interested in preference change occurring when new preference information, denoted by o, becomes available and has to be taken at face value, thereby prompting a change in prior preference information, denoted by π. The change, we require, should preserve as much useful information from π as can be afforded.

We believe that preference change thus described is a pervasive phenomenon, arising in many contexts straddling the realms of both human and artificial agency. Thus, there is a distinguished tradition in economics and philosophy puzzling over examples of conflict between an agent's subjective preference (what we call here the initial, or prior preference π) and a second-order preference, often standing for a commitment or moral rule (what we call here the new preference information o): the need to reconcile such a conflict is widely acknowledged, but in the absence of a concrete mechanism for doing so its presence is more often than not simply signaled as a problem that besets real life agents.

One example of the phenomenon that leads to preference revision occurs in the philosopher Harry Frankfurt's discussion on *second-order volitions*, which are desires about desires. According to Frankfurt, an agent's second-order volition with respect to a desire D is a desire that D be the agent's will, i.e., that D determines the agent's actions:

> Besides wanting and choosing and being moved to do this or that, [people] may also want to have (or not to have) certain desires and motives. They are

> capable of wanting to be different, in their preferences and purposes, from what they are. [Frankfurt, 1988]

The existence of second-order volitions is taken by Frankfurt to be a hallmark of personhood:

> ... it is having second-order volitions [...] that I regard as essential to being a person. [Frankfurt, 1988]

Frankfurt does not mention preferences explicitly but, since in standard models of rational agency preferences are taken to be part of what determines action, we can imagine that his point applies to preferences as well as desires. Preferences do appear in subsequent discussions, for instance in Robert Nozick's account of the same phenomenon:

> A person lacks rational integration when [they] prefer some alternative x to another alternative y, yet [they] prefer that [they] did not have this preference, that is, when [they] also prefer not preferring x to y to preferring x to y. When such a second-order preference conflicts with a first-order one, it is an open question which of these preferences should be changed. What is clear is that they do not hang together well, and a rational person would prefer that this not (continue to) be the case. [Nozick, 1994]

Nozick's formulation brings us closer to the framework we will be using in this chapter, of preference orders battling it out. If we assimilate his second order preferences to Frankfurt's second order volitions, then we have the same problem of a mismatch, or 'lack of integration', between the agent's preferences. Nozick confines himself to just labeling the problem and does not tell us how to solve it, but we are led to understand that it is, notwithstanding, a problem. The same issue arises in economist John C. Harsanyi's distinction between an individual's *social welfare function* and their *personal utility function*, two notions that refer to two distinct types of preferences:

> ... the former [type of preference] must express what this individual prefers (or rather, would prefer) on the basis of impersonal social considerations alone, and the latter must express what he actually prefers, whether on the basis of his personal interests or on any other basis. The former may be called his 'ethical' preferences, the latter his 'subjective' preferences. [Harsanyi, 1955]

In Harsanyi's phrasing, the problem is, as for Nozick, one of conflict between preferences: humans are the kind of beings that can entertain two types of preferences at the same time, and sometimes these preferences are not perfectly aligned. At the same time, Harsanyi seems to suggest, the ethical preferences (which we may assimilate with Frankfurt's second

order desires and Nozick's second order preferences) enjoy a certain type of prestige over the other, such that in a perfect world the subjective preferences would coincide with the ethical preferences, but in the imperfect world we live in the two are sometimes at odds with one another. That this is eminently possible is, of course, an old adage, and the phenomenon is sometimes called *akrasia* from the Greek term for lack of will. A classical example of akrasia occurs with habits that people want to get rid of, such as smoking: an agent smokes, which according to the revealed preference paradigm we have been espousing in the previous chapters, can be taken to indicate a preference of smoking over non-smoking; but at the same time, the agent might want to quit, i.e., to have the opposite preference. In the context of decision theory, this problem shows up in the form of Richard C. Jeffrey's Akrates, who would rather be abstinent than smoke:

> ...although Akrates cannot simply choose to prefer abstinence on this occasion—and that, after all, is why his preference for preferring abstinence is (uneasily) compatible with his preference for smoking—he can undertake a project of modifying his preferences over time, so that one day he may regularly prefer abstinence, just as now he regularly prefers smoking. The steps toward this desired end may involve hypnosis, reading medical textbooks, discussing matters with like-minded friends, or whatever. But in accounting for Akrates's undertaking of these activities it seems natural to cite his preference for preferring abstinence, just as in accounting for his activities as he flings drawers open and searches through pockets of suits, one may cite his preference for smoking... [Jeffrey, 1974]

Jeffrey, just like Frankfurt, Nozick and Harsanyi, is interested here only in pointing out the basic fact that Akrates can prefer smoking to abstinence, while wishing he had the opposite preference. Jeffrey assumes that Akrates' problem will be solved if Akrates can only get himself, by whatever means necessary, to change his disposition such that he comes to prefers abstinence to smoking. The challenge Akrates is facing is that he needs to make sure that his priorities are aligned with the preferences specific to a respected source. The same challenge can occur in technological applications, from updating CP-nets [Cadilhac et al., 2015] to changing the order in which search results are displayed on a page in response to user provided specifications. Similar topics are emerging in the discussion on ethical decision making for artificial agents [Rossi and Mattei, 2019]. Thus, far from being an issue of narrow interest, the problem of changing preferences and resolving conflicts along the way is an important component of rational agency.

Whether it is the internal conflict between an agent's private leanings and the better angels of its nature, or a content provider wanting to tailor its products for a better user experience, many cases of preference change involve a conflict between two types of preferences, one of which is perceived as having priority over the other: we will call this type of change *preference revision*, due to its similarity with revision as presented in Section 3.1.

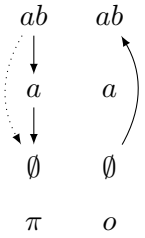


Figure 7.1: Preference order π has to be revised by preference o. A direct comparison ranking i better than j is depicted by a solid arrow from i to j, with comparisons inferred by transitivity depicted by dotted arrows. To distinguish these types of preference orders from the preference orders used to model belief change operators in assignments, we draw the better elements on top here.

Up to this point we have been arguing that cases of preference revision can arise in several contexts, but have not seen any actual mechanism for handling them. What is Akrates to do? Indeed, if there are only two alternatives to work with, e.g., smoking and abstinence, then it is difficult to see what more can be said, at the formal level, other than that smoking and abstinence should be swapped in Akrates' preference order. To see that there is more at stake here, we must look at an example with more than two alternatives.

Example 7.1

Consider the scenario from Example 1.6, where a doctor is considering treatments for a novel disease. The doctor has two drugs, a and b, that can be prescribed alone or in combination with each other. For the purposes of this illustration, let us assume the doctor is considering three alternatives: either to administer a and b together, denoted by the interpretation ab, or a alone, denoted by the interpretation a, or nothing, denoted by $\emptyset$.

The doctor's initial assessment is that a and b together work better than a alone, which by itself is better than doing nothing. This ranking constitutes a preference order, and by virtue of transitivity we can conclude that a and b together are better than administering nothing. We can write the doctor's preferences as the set $\pi = \{(ab, a), (a, \emptyset), (ab, \emptyset)\}$ of comparisons the doctor bases its assessment on. The comparison (ab, a), for instance, is to be read that ab is strictly better than a. Of the comparisons in π, we may assume that (ab, a) and $(a, \emptyset)$ are established by the doctor directly, while $(ab, \emptyset)$ is inferred by transitivity from the other two. The preference order π is depicted in Figure 7.1.

Afer a while, the doctor becomes convinced that $\emptyset$ is better than ab, and hence has to modify its initial assessment π accordingly. We can represent the new information as a preference order o in its own right. The preference order o consists of the

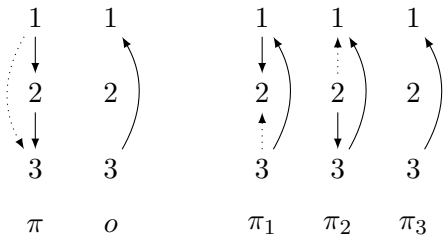


Figure 7.2: Revising preference order π by o: simply adding o to π leads to a cycle, so if o is accepted then a choice needs to be made regarding which of the initial comparisons of π to keep; potential candidates for the revised order are π_1, π_2 or π_3. A direct comparison ranking i better than j is depicted by a solid arrow from i to j, with comparisons inferred by transitivity depicted by dotted arrows.

comparison $(\emptyset, ab)$, i.e., $o = \{(\emptyset, ab)\}$, and is likewise depicted in Figure 7.1. Note that simply adding o to π leads to a cycle between the three alternatives, i.e., the transitive closure of $\pi \cup o$ implies that ab is strictly better than itself. Since the doctor is committed to accepting o, then they will have to give up something from π in order to maintain consistency.

Example 7.1 is not a case of akrasia, but illustrates the problem just as well: if new preference information contradicts an existing preference, in the sense that it leads to a preference cycle, then some of the comparisons involved in the cycle have to be given up. The challenge for Akrates, then, is that even after fitting the second order preference into his schedule, he still needs to figure out what the rest of his preference order looks like. In other words, Akrates has a decision problem on his hands: since there is no unique way, on logical grounds alone, of breaking a preference cycle, some extra information needs to be brought in. What should this extra information concern?

Based on our experience so far, we should expect the answer to be preferences: indeed, preferences on preferences themselves. This strategy, which we will be pursuing in the rest of this chapter, has already been anticipated by Amartya Sen:

> ...we need to consider rankings of preference rankings to express our moral judgments. [Sen, 1977]

We want to pick up Sen's suggestion and suggest that preferences over the basic building blocks of a preference order, i.e., the comparisons it is made of, offer a way of understanding preference revision.

Example 7.2

We revisit the scenarion in Example 7.1, with a doctor revising their preferences over three treatment options: ab, a and $\emptyset$. To further simplify the problem, we denote the alternatives by integers, such that 1 stands for ab, 2 stands for a and 3 stands for $\emptyset$. The initial preference π is such that, as a result of direct comparison, alternative 1 is ranked better than 2 and 2 is ranked better than 3; by virtue of transitivity, it is also inferred that 1 is considered better than 3. We want to revise π by a preference o, according to which 3 is better than 1. Both preference orders π and o are depicted in Figure 7.2.

The simplest solution is to add o to π (i.e., include the comparisons contained in both), but the transitivity requirement leads to a cycle between 1, 2 and 3, which we would like to avoid. We are thus in a situation where π and o cannot be jointly accepted, but since o, we stipulate, must be accepted, something must be given up from π (though, we ask, no more than strictly necessary). How is the decision to be made? We suggest that an implicit preference over the comparisons of π that were explicitly provided can provide an answer: if the comparison of 1-vs-2 (the edge from 1 to 2 in Figure 7.2) is preferred to the one of 2-vs-3 then the result is π_1, which holds on to 1-vs-2 from π and together with o infers, by transitivity, that 3 is better than 2; alternatively, a preference for 2-vs-3 over 1-vs-2 leads to π_2 as the result, while indifference between the two comparisons means that both are given up, resulting in π_3. Thus, preference over comparisons in π translates as choice over how to go about revising π. Interestingly, we may also reason in the opposite direction: observing choice behavior across different instances of revision allows us to infer preferences over comparisons in π, e.g., revising to π_1, rather than to π_2 or π_3, can be rationalized as saying that the comparison of 1-vs-2 is considered better than 2-vs-3.

Our aim in this chapter is to formalize the type of reasoning illustrated in Example 7.2 by rationalizing preference change as a type of choice function on what we will call the *direct comparisons of* π, i.e., the explicit preferences assumed to be given in π. Since a conflict between π and o forces some of the direct comparisons of π to be renounced, additional information in the form of a preference order over the direct comparisons of π will serve as guide to the choice function. Our purpose, in this, is not to legislate on what is the right choice to make; rather, it is to make sure that whatever the choice is, it is made in a coherent way. To this end, we present a set of rationality postulates to capture conditions under which the preference order on direct comparisons of π exists and has desired properties. Thus, the significance of our approach lies in laying bare the theoretical requirements and basic assumptions for mechanisms intended to revise preferences.

The postulates we put forward bear a distinct resemblance to the AGM postulates employed for belief revision [Alchourrón et al., 1985, Katsuno and Mendelzon, 1992, Hansson, 2017, Fermé and Hansson, 2018]: given that changing one's mind involves

choosing some parts of a belief to keep and some to remove, this is no coincidence. Indeed, the two problems are similar, though the structural particularities of preferences (in particular, the requirement that they are transitive) mean that transfer of insights from belief revision to preference revision is by no means straightforward.

7.1 Strict partial orders

We assume a finite set V of items, standing for the objects an agent can have preferences over. In keeping with the pattern established so far, a preference order is construed as a transitive binary relation on V, though in a break with standard practice the preferences that undergo change are denoted here by π and o, rather than $\leq$, to avoid confusion with the preferences used to model a belief change operator. Preferences that undergo revision are expected to satisfy, for any x, x_1 and x_2 in V, some combination of the following properties:

(Pr$_2$) If $x_1 \leq x_2$ and $x_2 \leq x_3$, then $x_1 \leq x_3$. (transitivity)

(Pr$_3$) If $x_1 \neq x_2$, then $x_1 \leq x_2$ or $x_2 \leq x_1$. (totality)

(Pr$_4$) $x \not\leq x$. (irreflexivity)

Properties Pr$_{2-4}$ accompany the properties provided in Section 2.2, which is why Pr$_1$ is absent from the current lineup. If π is a binary relation on a set V of items, then π is a *strict partial order (spo) on* V if π satisfies properties Pr$_2$ and Pr$_4$, i.e., if π is transitive and irreflexive. We write $\mathcal{O}_V$ for the set of strict partial orders on V. If π is an spo on a set V of items, then π is a *strict linear order on* X if π also satisfies property Pr$_3$, i.e., if π is total, in addition to being transitive and irreflexive. A *chain on* V is a strict linear order on a subset of V. We write $\mathcal{C}_V$ for the set of chains on V.

If π is an spo on a set of items V, then a *comparison* (i, j) *of* π is an element $(i, j) \in \pi$, for some items $i, j \in V$, interpreted as saying that, in the context of π, i is considered strictly better than j. To simplify notation, we sometimes also refer to comparisons with the letter c. We often have to consider the union $\pi_1 \cup \pi_2$ of two spos, which is not guaranteed to be an spo, since transitivity is not preserved under unions. If this is the case, we typically have to substitute $\pi_1 \cup \pi_2$ for its *transitive closure*, denoted by $(\pi_1 \cup \pi_2)^+$. Since preferences are required to be transitive, we write a sequence of comparisons $\{(1, 2), (2, 3) \ldots, (m{-}1, m)\}^+$ as $(1, \ldots, m)$.

If $\pi = (i_1, \ldots, i_m)$ is a chain on V, a *direct comparison of* π is a comparison $(i_k, i_{k+1}) \in \pi$, i.e., a comparison between i_k and its direct successor in π, with δ_π being the set of direct comparisons of π. The assumption is that direct comparisons are the result of explicit information, and are basic in the sense that they cannot be inferred by transitivity using other comparisons in π. Given preference orders $\pi \in \mathcal{C}_V$ and $o \in \mathcal{O}_V$, we want to

carve out the possible options for the revision of π by o. For this we use the set $\lfloor o \rfloor_\pi$ of *π-completions of o*, defined as:

$$\lfloor o \rfloor_\pi = \{(o \cup \delta)^+ \in \mathcal{O}_V \mid \delta \subseteq \delta_\pi\}.$$

The intuition is that a π-completion of o is a preference order constructed from o using some, and only, direct comparisons in π, i.e., information originating exclusively from the two sources given as input. We will expect that a preference revision operator selects one element of this set as the revision result.

Though taking $(\pi \cup o)^+$ as the result of revising π by o is not, in general, feasible, we still want to identify parts of $(\pi \cup o)^+$ that are uncontroversial. To that end, the *cycle-free part α^o_π of $(\pi \cup o)^+$* is defined as:

$$\alpha^o_\pi = \{(i, i{+}1) \in (\pi \cup o)^+ \mid (i{+}1, i) \notin (\pi \cup o)^+\},$$

i.e., the set of comparisons of $(\pi \cup o)^+$ not involved in a cycle with the comparisons of o. The *cyclic part κ^o_π of π with respect to o* is defined as:

$$\kappa^o_\pi = \{(i, i{+}1) \in \delta_\pi \mid (i{+}1, i) \in (\pi \cup o)^+\},$$

i.e., the set of direct comparisons of π involved in a cycle with o.

Example 7.3

For π and o as in Example 7.2, we have that $\delta_\pi = \{(1,2),(2,3)\}$, while the π-completions of o are $\lfloor o \rfloor_\pi = \{(3,1,2),(2,3,1),(3,1)\}$, i.e., the spos obtained by adding to o either of the elements of δ_π, or none (corresponding to π_1, π_2 and π_3). The cyclic part of π with respect to o is $\kappa^o_\pi = \delta_\pi = \{(1,2),(2,3)\}$ and the cycle-free part of π with respect to o is $\alpha^o_\pi = \emptyset$.

7.2 A general method for revising preferences

A *preference revision operator* $\triangleright$ is a function $\triangleright\colon \mathcal{C}_V \times \mathcal{O}_V \to \mathcal{O}_V$ taking a chain π and an spo o as input, and returning an spo $\pi \triangleright o$ as output. The choice of input and output can be motivated by a short nod to the material that is to come: since we will be rationalizing preference revision operators using preferences (i.e., preorders) on comparisons, an spo as output reflects the fact that certain comparisons are considered equally good, and must be given up together. The unfortunate effect of this, of course, is that the input and output formats do not match, which makes it impossible to iterate the revision operation. That being said, the output could be tightened to a chain: provided that the preferences guiding revision are a linear order (i.e., there are no ties); this will be touched on at the end of Section 7.4. Making both the input and output spos would be desirable, but intricacies of getting details right here means that this is best left for future work.

We start, then, by presenting a general procedure for revising preferences that, as advertised, utilizes total preorders on the set δ_π of direct comparisons of π: thus, a *preference assignment* $\preccurlyeq$ is a function $\preccurlyeq\colon \mathcal{C}_V \to \mathcal{T}_{V\times V}$ mapping every preference $\pi \in \mathcal{C}_V$ to a total preorder $\leq_\pi$ on elements of $V \times V$, i.e., on pairwise comparisons on the items of V, of which we are interested only in the preorder on δ_π. In typical AGM manner, a comparison $c_i \leq_\pi c_j$ in the context of a preorder $\leq_\pi$ on δ_π means that c_i is *better* than c_j.

If $\pi \in \mathcal{C}_V$, $o \in \mathcal{O}_V$ and $\leq_\pi$ is a total preorder on δ_π, then, for $i \geq 1$, the $\leq_\pi$-*level* i *of* δ_π, denoted $lvl^i_{\leq}(\delta_\pi)$, contains the i^{th} best elements of δ_π according to $\leq_\pi$, i.e., $lvl^1_{\leq_\pi}(\delta_\pi) = \min_{\leq_\pi}(\delta_\pi)$, $lvl^{i+1}_{\leq_\pi}(\delta_\pi) = \min_{\leq_\pi}(\delta_\pi \setminus \bigcup_{1\leq j\leq i} lvl^j_{\leq_\pi}(\delta_\pi))$, etc. Note that the $\leq_\pi$-levels of δ_π partition δ_π and, since δ_π is finite, there exists a $j > 0$ such that $lvl^i_{\leq_\pi}(\delta_\pi) = \emptyset$, for all $i \geq j$. The *addition operator* $\mathrm{add}^i_{\leq_\pi}(o)$ is defined, for any $o \in \mathcal{O}_V$ and $i \geq 0$, as follows:

$$\mathrm{add}^0_{\leq_\pi}(o) = (o \cup \alpha^o_\pi)^+,$$
$$\mathrm{add}^i_{\leq_\pi}(o) = \begin{cases} (\mathrm{add}^{i-1}_{\leq_\pi}(o) \cup (lvl^i_{\leq_\pi}(\delta_\pi) \cap \kappa^o_\pi))^+, & \text{if in } \mathcal{O}_V, \\ \mathrm{add}^{i-1}_{\leq_\pi}(o), & \text{otherwise.} \end{cases}$$

Intuitively, the addition operator starts by adding to o all the direct comparisons of π that are not involved in a cycle with it, i.e., which are not under contention by the accrual of new preference information. Then, at every further step $i > 0$, the addition operator tries to add all comparisons on level i of δ_π that are involved in a cycle with o: if the resulting set of comparisons can be construed as a spo (by taking its transitive closure) the operation is successful, and the new comparisons are added; if not, the addition operator does nothing. Since the addition of new comparison follows the order $\leq_\pi$, this ensures that better quality comparisons are considered before lower quality ones.

Note, this procedure guarantees that there are always *some* comparisons in $\pi \triangleright o$, i.e., $o \subseteq \pi \triangleright o$ regardless of anything else. Note, also, that the number of non-empty levels in δ_π is finite and the addition operation eventually reaches a fixed point, i.e., there exists $j \geq 0$ such that $\mathrm{add}^i_{\leq_\pi}(o) = \mathrm{add}^j_{\leq_\pi}(o)$, for any $i \geq j$. We denote by $\mathrm{add}^*_{\leq_\pi}(o)$ the fixed point of this operator and take it as the defining expression of a preference revision operator: if $\preccurlyeq$ is a preference assignment, then the $\preccurlyeq$-*induced preference revision operator* $\triangleright^{\preccurlyeq}$ is defined, for any $\pi \in \mathcal{C}_V$ and $o \in \mathcal{O}_V$, as:

$$\pi \triangleright^{\preccurlyeq} o \stackrel{\text{def}}{=} \mathrm{add}^*_{\leq_\pi}(o).$$

Note that, by design, $\mathrm{add}^*_{\leq_\pi}(o) \in \mathcal{O}_V$, i.e., the operator $\triangleright$ is well defined.

Example 7.4

For $\pi = (1, 2, 3, 4)$, $o = (3, 1)$, we obtain that $\delta_\pi = \{(1, 2), (2, 3), (3, 4)\}$. Suppose that there is a total preorder $\leq_\pi$ on δ_π according to which $(1, 2) <_\pi (2, 3) \approx_\pi (3, 4)$ (see Figure 7.3). To construct $\pi \triangleright o$, the addition operator starts from $\mathrm{add}^0_{\leq_\pi}(o) =$

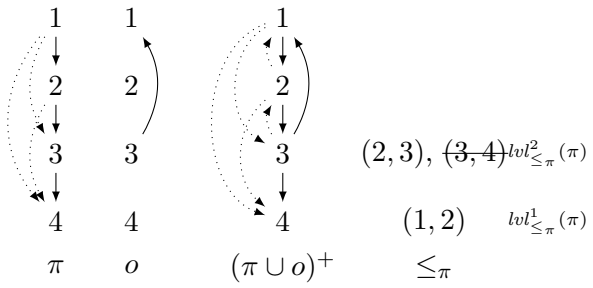


Figure 7.3: Preference revision by adding direct comparisons from π to o, using the preorder $\leq_\pi$. In $\leq_\pi$ lower means better; the comparison $(3,4)$ is ignored by the addition operator because it is not involved in a cycle with o (and is added at the beginning anyway).

$(\{(3,1)\} \cup \{(1,4),(2,4),(3,4)\})^+$, i.e., o and α^o_π. At the next step it tries to add $(1,2)$, which it can do successfully; at the next step is adds $(2,3)$, after which it runs out of comparisons to add.

7.3 Postulates

We show now that the procedure described in Section 7.2 can be characterized with a set of AGM-like postulates that do not reference any concrete revision procedure and are, by themselves, intuitive enough to provide reasonable constraints on any preference revision operator. The first two postulates apply to any $\pi \in \mathcal{C}_V$, $o \in \mathcal{O}_V$ and preference revision operator $\triangleright\colon \mathcal{C}_V \times \mathcal{O}_V \to \mathcal{O}_V$, and are as follows:

(P_1) $\pi \triangleright o \in \lfloor o \rfloor_\pi$.

(P_2) $\alpha^o_\pi \subseteq \pi \triangleright o$.

Postulates P_{1-2} are meant to capture preference revision in its most uncontroversial aspects, yet they still require some careful unpacking. Postulate P_1 states that $\pi \triangleright o$ is a π-completion of o, i.e., a preference order constructed only by adding direct comparisons from π to o, and, among other things, ensures that (i) $\pi \triangleright o \in \mathcal{O}_V$, (ii) $o \subseteq \pi \triangleright o$, and (iii) $\pi \triangleright o \subseteq (\pi \cup o)^+$. In terms of AGM propositional belief change, postulate P_1 does the same duty as the revision postulates R_1 and R_3 in Section 3.1, i.e., it sets limits for the revision result. However, the closest analogue to postulates P_{1-2} are enforcement postulates E_1 and E_3, respectively, in that they require the result to be formed by adding elements to the new information, and by requiring the result to be of a certain admissible type (refutable in enforcement, an spo here). Given this, a question emerges as to why not take condition (i)-(iii) as postulates instead of P_1: the reason is that, by requiring

$\pi \triangleright o$ to be constructed using only direct comparisons of π (in addition to o), postulate P_1 prevents $\pi \triangleright o$ from having opinions on items over which it had no opinions before, as illustrated in Example 7.5.

Example 7.5

For π and o as in Example 7.2, note that $\pi_4 = \{(3,1),(3,2)\}$ is such that $o \subseteq \pi_4 \subseteq (\pi \cup o)^+$. However, the comparison $(3,2)$ occurs neither in π nor in o as a direct comparison, and is entirely unjustified. By contrast, $(3,2)$ in $\pi_1 = (3,1,2)$ occurs as the result of inference from $(3,1)$, which is added from o, and $(1,2)$, which is preserved from π.

Postulate P_2 says that the cycle-free part of π with respect to o is to be preserved in $\pi \triangleright o$, and is meant to preserve the parts of $(\pi \cup o)^+$ that are not up for dispute. Note that in the case when $(\pi \cup o)^+$ does not contain a cycle then $\alpha^o_\pi = (\pi \cup o)^+$, and P_2 together with P_1 imply that $\pi \triangleright o = (\pi \cup o)^+$: this is the case when revision is easy, and nothing special needs to be done. In this, postulate P_2 serves the same function as the revision postulate R_2, but comes closest to enforcement postulate E_2, in the ideal case, when o can simply be added to π, results in the union of the two structures.

So far we have established that, if there is no conflict between π and o, then we can simply add o to π; and if there is a conflict, then $\triangleright$ must choose between the direct comparisons of π involved in the cycle. This choice, however, must be coherent, in a very precise sense, illustrated by Example 7.6.

Example 7.6

Consider revising $\pi = (1,2,3,4)$ from Example 7.4 by $o_1 = (4,1)$. This requires a choice between comparisons $(1,2)$, $(2,3)$ and $(3,4)$: assume $(1,2)$ is chosen, suggesting $(1,2)$ is better than $(2,3)$ and $(3,4)$. Suppose, now, that we revise π by $o_2 = \{(3,1)\}$. This requires a choice between $(1,2)$ and $(2,3)$: in accordance with the previous decision, $(1,2)$ should be chosen here as well.

The choice has to reflect an implicit preference order over the direct comparisons of π, and this is handled by the following postulates, meant to apply to $\pi \in \mathcal{C}_V$, $o_1, o_2 \in \mathcal{O}_V$ such that $(o_1 \cup o_2)^+ \in \mathcal{O}_V$, and a preference revision operator $\triangleright$:

(P_3) $\pi \triangleright (o_1 \cup o_2)^+ \subseteq ((\pi \triangleright o_1) \cup o_2)^+$.

(P_4) If $((\pi \triangleright o_1) \cup o_2)^+ \in \mathcal{O}_V$, then $((\pi \triangleright o_1) \cup o_2)^+ \subseteq \pi \triangleright (o_1 \cup o_2)^+$.

There is a similarity between postulates P_{3-4} and the revision postulates R_5 and R_6 from Section 3.1, but the parallel is closest to enforcement postulates E_{5-6} from Section 3.3. These postulates ensure that the choice between two options is stable and independent

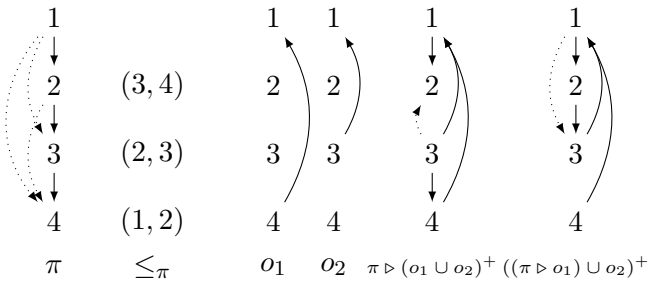


Figure 7.4: Postulates P_{3-4} are satisfied only if o_1 and o_2 are coordinated with respect to π.

of alternatives not directly involved. Postulates P_{3-4} are meant to ensure the same here: however, it turns out that in the present context this happens only under a specific set of conditions.

If o_1 and o_2 are spos, o_1 and o_2 are *coordinated with respect to* π if for any $\delta \subseteq \kappa_\pi^{o_1}$ such that for every direct comparison $(i, i{+}1) \in \delta$, neither $(i, i{+}1)$ nor $(i{+}1, i)$ is in $(o_1 \cup o_2)^+$, it holds that if $(o_1 \cup \delta)^+ \in \mathcal{O}_V$, then $((o_1 \cup o_2)^+ \cup \delta)^+ \in \mathcal{O}_V$. In other words, if π and o_1 form a cycle and we want to add o_2 as well, then we look at the direct comparisons in π that are not directly ruled out by $(o_1 \cup o_2)^+$, i.e., such that neitherm them nor their inverses are contained in $(o_1 \cup o_2)^+$. The property of coordination says that if we can consistently add some of these comparisons to o_1, then we can also add them to $(o_1 \cup o_2)^+$. Intuitively, coordination means that adding extra information o_2 does not step on o_1's toes by rendering unviable any comparisons that were previously viable. The following example makes this clearer.

Example 7.7

Take $\pi = (1, 2, 3, 4)$ and $o_1 = (4, 1)$, $o_2 = (3, 1)$. The direct comparisons of π that are involved in a cycle with o_1 are $\kappa_\pi^{o_1} = \{(1, 2), (2, 3), (3, 4)\}$, so that revision by o_1 requires making a choice between $(1, 2)$, $(2, 3)$ and $(3, 4)$. Notice that neither of $(1, 2)$, $(2, 3)$ and $(3, 4)$ is directly ruled out by $(o_1 \cup o_2)^+$: we have, for instance, that $(1, 2) \notin (o_1 \cup o_2)^+$ and $(2, 1) \notin (o_1 \cup o_2)^+$, and the same holds for $(2, 3)$ and $(3, 4)$. The significance of this is that adding o_2 to o_1 still makes the choice over which comparisons to keep be between $(1, 2)$, $(2, 3)$ and $(3, 4)$.

However, consider the set $\delta = \{(1, 2), (2, 3)\}$. We have that $(o_1 \cup \delta)^+ \in \mathcal{O}_V$, but $((o_1 \cup o_2)^+ \cup \delta)^+ \notin \mathcal{O}_V$, meaning that o_1 and o_2 are not coordinated with respect to π. In other words, whereas with o_1 we can add $(1, 2)$ and $(2, 3)$ together, with o_1 and o_2 we cannot add them anymore. This, then, makes it possible to add $(3, 4)$, irrespective of where it is in the preorder on comparisons.

At the same time, for the preorder $\leq_\pi$ in Figure 7.4 and the revision operator $\triangleright$

induced by it, we have that $(3,4) \in \pi \triangleright (o_1 \cup o_2)^+$, but $(3,4) \notin ((\pi \triangleright o_1) \cup o_2)^+$, i.e., postulate P_3 is not satisfied. The two facts are related, as the addition of o_2 tampers with the choice problem: though we can still add either one of the three comparisons, as mentioned above, we cannot add $(1,2)$ and $(2,3)$ together anymore, which in turn means that $(3,4)$ can be added regardless of its position in $\leq_\pi$.

The significance of coordination, as the following theorem shows, is that it is needed in order for postulates P_{3-4} to be effective at ensuring that choice across different types of incoming preferences is coherent.

Theorem 7.1

If $\preccurlyeq\colon \mathcal{C}_V \to \mathcal{T}_{V\times V}$ is a preference assignment and $\triangleright^{\preccurlyeq}$ is the $\preccurlyeq$-induced revision operator, then, $\triangleright^{\preccurlyeq}$ satisfies postulates P_{3-4} if and only if, for any $\pi \in \mathcal{C}_V$ and $o_1, o_2 \in \mathcal{O}_V$, it holds that o_1 and o_2 are coordinated with respect to π.

Proof

("$\Leftarrow$") Take $o_1, o_2 \in \mathcal{O}_V$ that are coordinated with respect to π. We will show that, for any preorder $\leq_\pi$ on δ_π, the $\preccurlyeq$-induced revision operator $\triangleright^{\preccurlyeq}$ satisfies postulates P_{3-4}. Since $\triangleright^{\preccurlyeq}$ satisfies postulates P_{3-4} trivially if $(\pi \cup o_1)^+ \in \mathcal{O}_V$, we look at the case when $\kappa_\pi^{o_1} \neq \emptyset$, i.e., when $(\pi \cup o_1)^+$ contains a cycle.

For postulate P_3, assume there is a comparison $c^\star \in \mathrm{add}^*_{\leq_\pi}(o_1 \cup o_2)^+$ such that $c^\star \notin (\mathrm{add}^*_{\leq_\pi}(o_1) \cup o_2)^+$. If $c^\star \in (o_1 \cup o_2)^+$ then a contradiction follows immediately. We thus have to look at the case when $c^\star \notin (o_1 \cup o_2)^+$, which contains two subcases of its own.

Case 1. If $c^\star \in \delta_\pi$, then by our assumption we have that $c^\star \in \kappa_\pi^{o_1}$, i.e., $c^\star$ is involved in some cycle with o_1. From $c^\star \notin \mathrm{add}^*_{\leq_\pi}(o_1)$ we infer that there must be a set $\delta \subseteq \delta_\pi$ of direct comparisons of π that precede $c^\star$ in $\leq_\pi$, are added to o_1 before it, and prevent $c^\star$ itself from being added. In particular, this means that $(o_1 \cup \delta)^+ \in \mathcal{O}_V$, but $((o_1 \cup \delta)^+ \cup \{c^\star\})^+ \notin \mathcal{O}_V$. At the same time, we know that $c^\star \in \mathrm{add}^*_{\leq_\pi}(o_1 \cup o_2)^+$, i.e., $c^\star$ can be consistently added to $(o_1 \cup o_2)^+$. Note that this happens after all the comparisons in δ, which precede it in $\leq_\pi$, have been considered as well. This implies that not all of the comparisons in δ can be added to $(o_1 \cup o_2)^+$, since if they could, then the cycle formed with o_1, δ and $c^\star$ would be reproduced here as well. If not all of the comparisons in δ can be added to $(o_1 \cup o_2)^+$, this must be because $((o_1 \cup o_2)^+ \cup \delta)^+$ contains a cycle, i.e., $((o_1 \cup o_2)^+ \cup \delta)^+ \notin \mathcal{O}_V$. This now contradicts the fact that o_1 and o_2 are coordinated with respect to π.

Case 2. If $c^\star$ is not a direct comparison of π, then it is inferred by transitivity using at least one direct comparison of π added previously. We apply the reasoning in

Case 1 to these direct comparisons to show that they are in $(\text{add}^*_{\leq_\pi}(o_1) \cup o_2)^+$, which implies the conclusion as well.

For postulate P_4, take $c^\star \in (\text{add}^*_{\leq_\pi}(o_1) \cup o_2)^+$ and assume $c^\star \notin \text{add}^*_{\leq_\pi}(o_1 \cup o_2)^+$. As before, the non-obvious case is when $c^\star \notin (o_1 \cup o_2)^+$. If $c^\star \in \delta_\pi$, then from the assumption that $c^\star \notin \text{add}^*_{\leq_\pi}(o_1 \cup o_2)^+$ we conclude that there is a set $\delta \subseteq \kappa_\pi^{o_1}$ of comparisons that precede $c^\star$ in $\leq_\pi$, are added to $(o_1 \cup o_2)^+$ before it and, in concert with $(o_1 \cup o_2)^+$, block $c^\star$ from being added, i.e., $((o_1 \cup o_2)^+ \cup \delta)^+ \in \mathcal{O}_V$ but $((o_1 \cup o_2)^+ \cup \delta')^+ \notin \mathcal{O}_V$, where $\delta' = \delta \cup \{c_\star\}$. From the second to last result we infer that δ can be added consistently to $(o_1 \cup o_2)^+$ and, since we have that $c^\star \in (\text{add}^*_{\leq_\pi}(o_1) \cup o_2)^+$ as well, we obtain that and $c^\star$ can be added consistently to o_1. In other words, it holds that $(o_1 \cup \delta')^+ \in \mathcal{O}_V$, which contradicts the coordination assumption. The case when $c^\star \notin (o_1 \cup o_2)^+$ is treated analogously as for postulate P_3.

("$\Rightarrow$") Assume that there are $o_1, o_2 \in \mathcal{O}_V$ not coordinated with respect to π, i.e., there exists a set $\delta \subseteq \kappa_\pi^{o_1}$ of direct comparisons of π that are involved in a cycle with o_1 and are such that $(o_1 \cup \delta)^+ \in \mathcal{O}_V$ and $((o_1 \cup o_2)^+ \cup \delta)^+ \notin \mathcal{O}_V$. Additionally, we have that neither of the comparisons in δ, or their inverses, are in $(o_1 \cup o_2)^+$. We infer that there must exist a comparison $c^\star \in (\kappa_\pi^{o_1} \setminus \delta)$ that completes the cycle. We will show that there exists a preorder $\leq_\pi$ such that the revision operator induced by it does not satisfy P_3. Take a preorder $\leq_\pi$ on δ_π that arranges the elements of δ in a linear order at the bottom of $\leq_\pi$, i.e., such that $c_j <_\pi c_l$, for any $c_j \in \delta$ and $c_l \notin \delta$, and $c^\star$ the maximal element in $\leq_\pi$, i.e., $c_j <_\pi c^\star$, for any $c_j \in \delta$. This implies, in particular, that $c_j <_\pi c^\star$, for any $c_j \in \delta$. Note, now, that $c^\star \in \text{add}^*_{\leq_\pi}(o_1 \cup o_2)^+$: this is because, by assumption, not all of the comparisons in δ can be added to $(o_1 \cup o_2)^+$, and this makes it possible for $c^\star$ to be added. On the other hand, $c^\star \notin (\text{add}^*_{\leq_\pi}(o_1) \cup o_2)^+$: this is because here we can, again by assumption, consistently add δ to o_1 and, since $c^\star$ is the last in line to be added, the inevitability of creating a cycle with δ and the rest of the comparisons of o_1 makes it impossible to do so consistently. We obtain that $c^\star \in \text{add}^*_{\leq_\pi}(o_1 \cup o_2)^+$ but $c^\star \notin (\text{add}^*_{\leq_\pi}(o_1) \cup o_2)^+$, i.e., postulate P_3 is not satisfied. Concurrently, there will be a comparison in δ that occurs in $(\text{add}^*_{\leq_\pi}(o_1) \cup o_2)^+$ that does not make it into $\text{add}^*_{\leq_\pi}(o_1 \cup o_2)^+$, showing that P_4 is not satisfied either.

Theorem 7.1 shows that coordination is needed in order to make sure that postulates P_{3-4} work, and we will henceforth assume that o_1 and o_2 are coordinated with respect to π whenever we apply these postulates.

7.4 Preference revision as choice over comparisons

We show now that the procedure described in Section 7.2 is characterized by the postulates introduced in Section 7.3, under the restrictions established through Theorem 7.1. Theorem 7.2 shows that the procedure in Section 7.2 yields preference revision operators that satisfy postulates P_{1-4}.

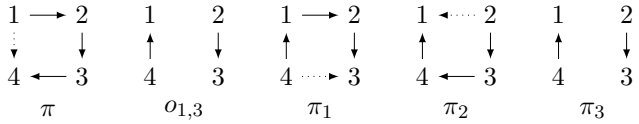


Figure 7.5: Revision of π by $o_{1,3}$ forces a choice between direct comparisons $(1,2)$ and $(3,4)$: since keeping both $(1,2)$ and $(3,4)$ is not possible, at least one of them, potentially both, must be discarded. Depending on the choice made, possible results are π_1, π_2 and π_3.

Theorem 7.2

If $\preccurlyeq\colon \mathcal{C}_V \to \mathcal{T}_{V\times V}$ is a preference assignment, then the revision operator $\triangleright^{\preccurlyeq}$ induced by it satisfies postulates P_{1-4}, for any $\pi \in \mathcal{C}_V$ and $o, o_1, o_2 \in \mathcal{O}_V$ such that o_1, o_2 are coordinated with respect to π.

Proof

Satisfaction of postulates P_{1-2} is straightforward. For P_1, since at every step $\mathrm{add}^i_{\leq_\pi}$ selects some direct comparisons in π to add to o, the end result satisfies the condition for being in $\lfloor o \rfloor_\pi$. For P_2, note that $c(\pi \cup o)^+ \subseteq \mathrm{add}^0_{\leq_\pi}(o) \subseteq \mathrm{add}^*_{\leq_\pi}(o)$. Since o_1 and o_2 are assumed to be coordinated with respect to π, satisfaction of postulates P_{3-4} is guaranteed by Theorem 7.1.

For the converse, we want to show that any preference revision operator satisfying P_{1-4} can be rationalized using a preference assignment. To that end, we will construct the preorder $\leq_\pi$ from binary comparisons, but we must first figure out how to compare two direct comparisons $(k, k{+}1)$ and $(l, l{+}1)$. This is done by creating a situation where we cannot add both and hence one has to be given up. We will use a special type of preference order to induce a choice between these comparisons. If $\pi \in \mathcal{C}_V$ and $(k, k{+}1), (l, l{+}1) \in \delta_\pi$, the *choice inducing preference* $o_{k,l}$ *for* $(k, k{+}1)$ *and* $(l, l{+}1)$ is defined as $o_{k,l} = \{(k{+}1, l), (l{+}1, k)\}$.

Example 7.8

To induce a choice between direct comparisons $(1,2)$ and $(3,4)$ in Figure 7.5, revise by $o_{1,3} = \{(2,3),(4,1)\}$. Note that effectiveness of this maneuver hinges on the choice being confined to the direct comparisons of π: if inferred comparisons were allowed to be part of the choice, $o_{1,3}$ loses its power to discriminate between $(1,2)$ and $(3,4)$: if, for instance, $(1,3)$ and $(2,4)$ are chosen, then $(2,1)$ and $(4,3)$ have to be inferred, leaving no space for a choice between $(1,2)$ and $(3,4)$, i.e., $o_{1,3}$ would tell us nothing about the implicit preference between $(1,2)$ and $(3,4)$. We can also see

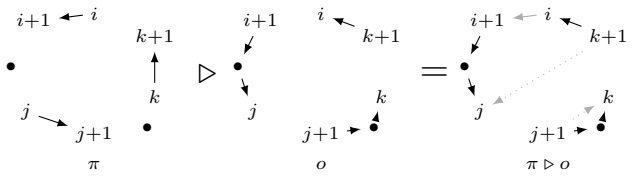


Figure 7.6: To show that $\leq_\pi^\triangleright$ is transitive, we show first that $(k, k{+}1) \notin \pi \triangleright o$. Bullets indicate other potential items in π; faded arrows indicate comparisons that may not be in $\pi \triangleright o$, but can be consistently added to it.

that comparison of $(1, 2)$ and $(2, 3)$ is done by revising by $(3, 1)$.

If $(k, k{+}1), (l, l{+}1) \in \delta_\pi$ and $\triangleright$ is a preference revision operator, then the *revealed order* $\leq_\pi^\triangleright$ *between* $(k, k{+}1)$ *and* $(l, l{+}1)$ is defined as:

$$(k, k{+}1) \leq_\pi^\triangleright (l, l{+}1) \text{ if } (l, l{+}1) \notin \pi \triangleright o_{k,l}.$$

Intuitively, $(l, l{+}1)$ being discarded from $\pi \triangleright o_{k,l}$ signals that it is considered less important than $(k, k{+}1)$.

Lemma 7.1

If $\triangleright$ satisfies postulates P_{1-4}, then the revealed preference relation $\leq_\pi^\triangleright$ is transitive.

Proof

Take $\pi \in \mathcal{C}_V$ and $(i, i{+}1), (j, j{+}1), (k, k{+}1) \in \delta_\pi$ such that $(i, i{+}1) \leq_\pi^\triangleright (j, j{+}1) \leq_\pi^\triangleright (k, k{+}1)$ (we can assume that $i < j < k$). To show that $(i, i{+}1) \leq_\pi^\triangleright (j, j{+}1)$, take $o \in \mathcal{O}_V$ that contains all direct comparisons in π up to k, except $(i, i{+}1)$, $(j, j{+}1)$ and $(k, k{+}1)$, plus the comparison $(k{+}1, i)$. In other words, o is such that if $(i, i{+}1)$, $(j, j{+}1)$ and $(k, k{+}1)$ were added to it, a cycle would form. The first step involves showing that $(k, k{+}1) \notin \pi \triangleright o$. To see why this is the case, note first that, by design, not all of $(i, i{+}1)$, $(j, j{+}1)$ and $(k, k{+}1)$ can be in $\pi \triangleright o$, i.e., at least one of them must be left out. We now do a case analysis to show that, either way, $(k, k{+}1)$ ends up being left out.

Case 1. If $(k, k{+}1) \notin \pi \triangleright o$, the conclusion is immediate.

Case 2. If $(j, j{+}1) \notin \pi \triangleright o$, then we can safely add $(i, i{+}1)$ to $\pi \triangleright o$: this is because the inference of the opposite comparison, i.e., $(i{+}1, i)$, can be done only by adding all comparisons on the path from $i{+}1$ to i, and the absence of $(j, j{+}1)$ means this inference is blocked. Using P_{3-4} we can now conclude that $((\pi \triangleright o) \cup \{(i, i{+}1)\})^+ = \pi \triangleright (o \cup \{(i, i{+}1)\})^+$ (see Figure 7.6). Note, we can separate $o \cup \{(i, i{+}1)\}$ into $o_{j,k} = \{(k{+}1, j), (j{+}1, k)\}$ and all the comparisons on the path from $k{+}1$ to j, plus

the comparisons on the path from $j{+}1$ to k. Call this latter preference o'. We thus have that $(o \cup \{(i, i{+}1)\})^+ = (o_{j,k} \cup o')^+$ and, applying P_3, we obtain that:

$$\pi \triangleright (o \cup \{(i, i{+}1)\})^+ = \pi \triangleright (o_{j,k} \cup o')^+ \subseteq ((\pi \triangleright o_{j,k}) \cup o')^+.$$

Since, by definition, $(k, k{+}1) \notin \pi \triangleright o_{j,k}$ and $(k, k{+}1) \notin o'$, It follows that $(k, k{+}1) \notin \pi \triangleright (o \cup \{(i, i{+}1)\})^+$, then $(k, k{+}1) \notin ((\pi \triangleright o) \cup \{(i, i{+}1)\})^+$, and, finally, that $(k, k{+}1) \notin \pi \triangleright o$.

Case 3. If $(i, i{+}1) \notin \pi \triangleright o$, then we can safely add $(k, k{+}1)$ to $\pi \triangleright o$ and, by reasoning similar to above, show that $(j, j{+}1) \notin \pi \triangleright o$. Here we invoke Case 2.

With the fact that $(k, k{+}1) \notin \pi \triangleright o$ in hand, we can add $(j, j{+}1)$ to $\pi \triangleright o$ (by reasoning similar to above), because the path from $j{+}1$ to j in $\pi \triangleright o$ is blocked by the absence of $(k, k{+}1)$. Using postulates P_{3-4}, we conclude that:

$$\begin{aligned}((\pi \triangleright o) \cup \{(j, j{+}1)\})^+ &= \pi \triangleright (o \cup \{(j, j{+}1)\})^+ \\ &= \pi \triangleright (\{(i{+}1, \ldots, k), (k{+}1, i)\})^+ \\ &= ((\pi \triangleright o_{i,k}) \cup \{(i{+}1, \ldots, k)\})^+.\end{aligned}$$

Since $(k, k{+}1) \notin ((\pi \triangleright o) \cup \{(j, j{+}1)\})^+$, we conclude that $(k, k{+}1) \notin \pi \triangleright o_{i,k}$, which implies that $(i, i{+}1) \leq^{\triangleright}_{\pi} (k, k{+}1)$.

Lemma 7.1 is crucial for the following representation result.

Theorem 7.3

If $\triangleright$ is a revision operator satisfying postulates P_{1-4}, for any $\pi \in \mathcal{C}_V$ and $o, o_1, o_2 \in \mathcal{O}_V$ such that o_1, o_2 are coordinated with respect to π, then there exists a preference assignment $\preccurlyeq$ such that $\triangleright$ is the $\preccurlyeq$-induced revision operator.

Proof

For any $\pi \in \mathcal{C}_V$, take $\leq_\pi$ to be the revealed preference relation $\leq^{\triangleright}_{\pi}$. By Lemma 7.1, we know that $\leq_\pi$ is transitive, so the only thing left to is show is that $\pi \triangleright o = \mathrm{add}^*_{\leq_\pi}(o)$. We do this in two steps.

("$\subseteq$") For one direction, Take $(j, k) \in \pi \triangleright o$ and suppose $(j, k) \notin \mathrm{add}^*_{\leq_\pi}(o)$. Clearly, it cannot be the case that $(j, k) \in o$, so we conclude that (j, k) is either a direct comparison of π, or is inferred by transitivity using direct comparisons in π and o.

Case 1. If $(j, k) \in \delta_\pi$, then we can write (j, k) as $(j, j{+}1)$, Suppose that $(j, j{+}1)$ is on level i of δ_π: this means that if $(j, j{+}1)$ does not get added to $\mathrm{add}^*_{\leq_\pi}(o)$ at step i, then, since it cannot be inferred by transitivity, it does not get added at all.

The fact that $(j, j{+}1) \notin \mathrm{add}^*_{\leq_\pi}(o)$ thus means that $(j, j{+}1)$ forms a cycle with some comparisons in o and comparisons in π on levels $l \leq i$. First, note that $(j, j{+}1)$ cannot form a cycle with elements of o only, since that would imply that $(j{+}1, j) \in o$ and that would exclude the possibility that $(j, j{+}1) \in \pi \triangleright o$. Thus, at least one other comparison in the cycle must come from π. We can state, now, that, since $(j{+}1, j) \in \pi \triangleright o$, then at least one of these comparisons must be absent in $\pi \triangleright o$, i.e., there exists a direct comparison $(k, k{+}1) \in \delta_\pi$ such that $(k, k{+}1) \in lvl^j_{\leq_\pi}(\pi)$, for some $j \leq i$, $(k, k{+}1) \notin \pi \triangleright o$ and $(j, j{+}1)$, $(k, k{+}1)$, plus some other comparisons in o and π form a cycle. This means that it is safe to add o' to $\pi \triangleright o$, where o' contains all comparisons on the path from $k{+}1$ to j, plus the comparison on the path from $j{+}1$ to k. We can rewrite o' by separating out $(k{+}1, j)$ and $(j{+}1, k)$, i.e., $o' = (o_{j,k} \cup o')^+$. Applying postulates P_{3-4}, we now get that

$$\begin{aligned}((\pi \triangleright o) \cup o')^+ &= \pi \triangleright (o \cup o')^+ \\ &= \pi \triangleright (o_{j,k} \cup o')^+ \\ &\subseteq ((\pi \triangleright o_{j,k}) \cup o')^+.\end{aligned}$$

Using the assumption that $(j, j{+}1) \in \pi \triangleright o$ and the fact that $(j, j{+}1) \notin o'$, we can thus infer that $(j, j{+}1) \in \pi \triangleright o_{j,k}$. This, in turn, implies that $(j, j{+}1) <_\pi (k, k{+}1)$ and hence $(j, j{+}1)$ belongs to a lower level of δ_π than $(k, k{+}1)$: but this contradicts the conclusion drawn earlier that $(k, k{+}1)$ belongs to a level $l \leq i$, where i is the level of $(j, j{+}1)$.

Case 2. If (j, k) is not a direct comparison of π, then it is inferred from some direct comparisons of π that end up in $\pi \triangleright o$, together with comparisons in o. We can now apply the reasoning from Case 1 to the direct comparisons of π that go into inferring (j, k), to show that they must be in $\mathrm{add}^*_{\leq_\pi}(o)$. This, in turn, implies that (j, k) will be in $\mathrm{add}^*_{\leq_\pi}(o)$ as well.

The reasoning for the other direction is similar.

Theorems 7.2 and 7.3 describe preference revision operators that rely on total preorders $\leq_\pi$ on δ_π, where a tie between two direct comparisons means that if they cannot both be added, then they are both passed over. We can eliminate this indeciseveness by using *linear* orders on δ_π instead of preorders: this ensures that any two direct comparisons of π can be clearly ranked with respect to each other, and that a revision operator is always in a position to choose among them. On the postulate site, linear orders can be characterized by tightening the notion of a π-completion and, with it, postulate P_1. Thus, a *decisive π-completion of o* is defined as:

$$\lfloor o \rfloor^{\mathsf{D}}_\pi \stackrel{\text{def}}{=} \{(o \cup \delta)^+ \in \mathcal{O}_V \mid \emptyset \subset \delta \subseteq \delta_\pi\}.$$

The decisive version of P_1 is written, for any $\pi \in \mathcal{C}_V$ and $o \in \mathcal{O}_V$, as:

$(\mathrm{P}_{\mathsf{D}})$ $\pi \triangleright o \in \lfloor o \rfloor^{\mathsf{D}}_\pi$.

A *decisive preference assignment* $\preccurlyeq$ is a function $\preccurlyeq\colon \mathcal{C}_V \to \mathcal{C}_{V\times V}$ mapping every $\pi \in \mathcal{C}_V$ to a linear preorder $<_\pi$ on δ_π. We can now show the following result.

Theorem 7.4

A revision operator $\triangleright$ satisfies postulates P_D and P_{2-4} if and only if there exists a decisive preference assignment a such that, for any $\pi \in \mathcal{C}_V$ and $o, o_1, o_2 \in \mathcal{O}_V$ such that o_1, o_2 are coordinated with respect to π, $\triangleright$ is the $\preccurlyeq$-induced preference revision operator.

Proof

The proofs for Theorems 7.2 and 7.3 work here with minimal adjustments. Note that when choosing between two direct comparisons, postulate P_D does not allow $\triangleright$ to be indifferent anymore. This means that the revealed preference relation on δ_π ends up being linear.

7.5 Concrete preference revision operators

Theorems 7.2, 7.3 and 7.4 articulate an important lesson: preference revision performed in a principled manner, i.e., in accordance with P_{1-4} or P_D and P_{2-4}, involves having preferences over comparisons. Thus, to obtain concrete operators one must look at ways of ranking the comparisons in a preference π. We present here two simple solutions.

The *trivial assignment* $\preccurlyeq^{\mathrm{t}}$ and the *lexicographic assignment* $\preccurlyeq^{\mathrm{lex}}$ are defined by taking $(i, i{+}1) \approx^{\mathrm{t}}_\pi (j, j{+}1)$ and, respectively, $(i, i{+}1) <^{\mathrm{lex}}_\pi (j, j{+}1)$ if $i < j$, for any $\pi \in \mathcal{C}_V$ and $(i, i{+}1), (j, j{+}1) \in \pi$, These assignments induce the *trivial* and *lexicographic* operators $\triangleright^{\mathrm{t}}$ and $\triangleright^{\mathrm{lex}}$, respectively. Note that $\leq^{\mathrm{t}}_\pi$ is a preorder and $<^{\mathrm{lex}}_\pi$ is a linear order, prompting the following result.

Proposition 7.1

The operators $\triangleright^{\mathrm{t}}$ and $\triangleright^{\mathrm{lex}}$ satisfy postulates P_{1-4} and P_D and P_{2-4}, respectively.

Example 7.9

For π and o as in Example 7.2, we obtain that $\pi \triangleright^{\mathrm{t}} o = (31)$ and $\pi \triangleright^{\mathrm{lex}} o = (312)$.

7.6 Related work

Our work comes on the heels of existing research, but manages to carve its own niche in an otherwise populated landscape. Contrary to some previous work labeled as revision

of preferences [Bradley, 2007, Lang and van der Torre, 2008, Liu, 2011], we do not look at changes in preferences elicited by a change in beliefs: not because this is not an interesting topic, but because we are more interested in the mechanism of preference change independently of any other cognitive operations running at the same time.

More involved work [Cadilhac et al., 2015] describe preference change when preferences are represented using CP-nets [Boutilier et al., 2004] or dynamic epistemic logic [Benthem and Liu, 2014]. In contrast, we have opted to represent preferences as strict partial orders over a set of items: we believe this straightforward formulation allows the basic issue signaled by Amartya Sen [Sen, 1977], and mentioned here in the beginning of the chapter, to be visible and tackled head on.

Apart from the examples presented here from economics and philosophy, the basic phenomenon of preference change has also been raised in explicit connection to belief change [Hansson, 1995, Grüne-Yanoff and Hansson, 2009, Grüne-Yanoff, 2013], but a representation in terms of preferences on the comparisons present in the preference orders, along the lines we presented here, was not given. Most existing work on the topic proceeds by putting forward some concrete way of inserting some new preference into an existing one, possibly by shifting some elements of the original preference around, and occasionally with a remark on the similarity between this operation and a belief revision operation [Freund, 2004, Chomicki and Song, 2005, Liu, 2011, Ma et al., 2012]. None of these models, however, provides an analysis in terms of postulates or representation results in the manner described here.

7.7 Conclusion

We have presented a model of preference change, according to which revising a preference π goes in hand with having preferences over the comparisons of π, thereby providing a rigorous formal treatment to intuitions found elsewhere in the literature [Sen, 1977, Grüne-Yanoff and Hansson, 2009]. Interestingly, the postulates describing preference revision are analogous to the postulates for propositional enforcement, presented in Section 3.3. Our treatment unearthed interesting aspects of preference revision, such as the issue of coordination between successive instances of new preference information (Section 7.3) and the non-obvious solution to the question of how to rank two comparisons relative to each other (Section 7.4). These aspects are taken for granted in regular propositional revision, but prove key to successful application of revision to the more specialized context of transitive relations on a set of items, i.e., preference orders. In this respect, preference revision is akin to revision for fragments of propositional logic [Delgrande et al., 2018], and raises the possibility of exporting this approach to other formalisms in this family. The addition procedure in particular, which is directly modeled from the addition operation for propositional enforcement, lends itself to application in other formalisms by slight tweaking of the acceptance condition, and could thus supply some interesting lessons for revision in general.

There is also ample space for future work with respect to the present framework itself.

To facilitate exposition of the main ideas we imposed certain restrictions on the primary notions. Lifting these restrictions would yield broader results that would potentially cover more ground and apply to a more diverse set of inputs. We can consider, for instance, revising strict partial orders in general (not just linear orders), and using rankings that involve all comparisons of the initial preference order (not just the direct ones). As the space of possibilities becomes larger, the choice problems on this space become increasingly more complex as well. Finding the right conditions under which the choice mechanism corresponds to a set of appealing postulates requires a delicate balance of many elements, and holds the promise for interesting results.

CHAPTER 8

Conclusions

In this final chapter we look back at the road traveled so far and gather some thoughts about how things fit together, and where to go from here. This chapter serves as a summary of the contents of the thesis, a reflection on its connections to other, more broadly related work, and a pointer to directions for future research.

In putting together the material for this thesis we took seriously a claim made by Hans Rott and others [Rott, 2001, Bonanno, 2009, Arló-Costa and Pedersen, 2010] that changing beliefs is like making a decision. According to this viewpoint, revision is analogous to a single agent making a decision as to what possible outcomes out of a given menu it will focus on, where the menu consists of the allowed outcomes provided by the new information; update is a variation on this, according to which the final decision is distributed across all the models of the prior information; and merging is analogous to a group of agents deciding on the collective set of acceptable outcomes, subject to a constraint. The parallel with decision making was facilitated by the fact that the revision postulates R_1, R_3 and R_{5-6}, as well as the update postulates U_1, U_3 and U_{5-6}, are close analogues of axioms C_{1-4} for individual rational choice, and that merging postulates M_{0-8}, besides rehashing the revision postulates, also closely track properties typically employed to characterize voting rules. This parallel, we argued in Chapters 1 and 3, also makes sense on a conceptual level: the preferences that lie at the heart of rational choice, individual as well as social, reappear in belief change as preorders over outcomes, encoding the agents' assessments of the plausibility, or desirability, of outcomes relative to each other.

More broadly, the idea that agents use something along the lines of preference information when drawing inferences in the wild fits with a distinct line of research on the way in which non-monotonic logics look like at the semantic level [Strasser and Antonelli, 2019]. The idea, simply put, is that when agents use their background information, which we may call φ, to figure out whether something, which we may call μ, holds in the real world, what they do is that they pick *some* models of φ on top of which to reason. What exactly

this picking represents has never been entirely settled, but we can readily see that, from a cognitive point of view, it makes eminent sense: if the agent had to consult all the models of φ before it could make up its mind as to whether μ follows from it, as classical logic instructs, then it would probably never reach any conclusion, since the number of possibilities is likely to be astronomically high; and even if the agent did manage to reach a conclusion in efficient time, the answer would probably be, more often than not, *no*, since most real world inferences do not account for all the subtle, but entirely irrelevant, ways in which a scenario can be varied. Rather, we can imagine that real world agents draw inferences by picking something like the most 'normal', 'typical', or 'probable' models of φ and checking those to see if μ holds in them. Of course, the agent does not literally go through a list and picks out models of φ: a specialized module of its cognitive apparatus, e.g., its memory, attention or social background, does this for it. Thus, it could be argued, ambitiously, that all of non-monotonic reasoning, in general, is about choice: choice over which of the myriad possible configurations of the world to use in a specific reasoning task. And we can picture the rational choice theorists of yore pointing out that this process can be described, as it actually has been [Shoham, 1987, Pearl, 1989, Kraus et al., 1990], using choice functions and preference orders.

In Sections 3.1, 3.2 and 3.4 we presented the formal models for revision, update and merging, respectively, in the light of this preference-driven, choice theoretic approach. In doing so, we merely retraced steps taken by our predecessors [Rott, 2001, Bonanno, 2009, Arló-Costa and Pedersen, 2010, Konieczny and Pino Pérez, 2011], steps that were present even in the original models of belief revision [Alchourrón et al., 1985, Katsuno and Mendelzon, 1992].

In Section 3.3 we showed that the choice theoretic perspective can also be useful for the design of new belief change operators, and exemplified this on *enforcement*, a dual version of revision that sits somewhere on the spectrum of non-prioritized belief change operators. The main challenge, for us, of figuring out what enforcement does was to understand it at the semantic level: what do the preorders look like? And what kind of choice function best fits enforcement? Originally, we opted for a representation in terms of partial orders on formulas, or sets of interpretations [Haret et al., 2018c], with the choice function picking out the one set that was best, given the new information: the partial order, then, had to be designed in such a way that there would always be a unique best set of interpretations out of any lineup that could be presented, and the specification of the conditions under which this held true ended up being rather opaque. In this work we switched to a more standard representation, in terms of preorders on the interpretations themselves; what had to be changed, then, was the choice function: we could not use something that selected models of μ, since the models of μ needed to be left in place. What we needed was a function that added models to μ in as greedy manner as possible, and this led us to the idea of the addition operator. The idea behind enforcement proved to be more fertile than we thought it would be, as it plied itself naturally to revision of preferences, described in Chapter 7. The original aim for enforcement, which was to provide a principled approach to enforcement in abstract argumentation [Baumann,

2012], ended up being sidelined, but is a promising direction for future work.

The same choice theoretic perspective, applied back to revision, led us to think about the role of the different postulates in the grand scheme of things. It became clear that postulate R_2 was not a rationality constraint in the same manner as the other postulates were, in the sense that it concerned exclusively the placement of the models of φ in the agent's ranking on outcomes, and corresponded to something like the agent's attitude, or bias, about how privileged these models should be when revision needed to occur: in this perspective, R_2 could be seen as one attitude among many. A more systematic attempt to generate such biases, using simple variations of the functions used to rank outcomes relative to φ, i.e., the *aggregation functions* in Section 2.3, led to Chapter 4. Of course, more sophisticated variations, corresponding to more psychologically realistic biases can be imagined, and it is an exciting prospect to think of revision along these parameters. At the same time, the more fine grained view on the types of biases an agent can have towards its initial beliefs raises the question of what these attitudes are good for, i.e., whether they can be used for tasks such as learning or tracking the truth [Kelly, 1998, Baltag et al., 2019]. The idea here is to view revision as part of an ongoing process by which the agent continuously refines its representation of the outside world, with the aim of settling on stable, correct information. Such a task, we think, provides a natural benchmark for revision operators, and it has the potential to connect belief revision to other topics of importance to the field of AI. It would also be interesting to study the complexity of these operators, and see how it compares to the complexity of existing belief change operators [Eiter and Gottlob, 1992, Pfandler et al., 2015].

In Chapter 5 we looked at merging, which is to revision as social choice is to individual rational choice. From the onset we opted to look at merging as a collective decision process, whose aim is to be fair, rather than as an information aggregation process, whose aim would be to be right, or accurate. Postulates M_{0-8} are, largely, compatible with both approaches. The idea of looking at merging as a kind of voting scenario, where the candidates are the outcomes, suggested that postulates M_{0-8} were only a starting point, and that merging was fair game for the large variety of properties studied in social choice. This led to the original paper [Haret et al., 2016b] and to Sections 5.1, 5.2 and 5.3, which are based on it. Shortly after, the *Handbook of Computational Social Choice* [Brandt et al., 2016] and the volume on *Trends in Computational Social Choice* [Endriss, 2017] came out, and it became clear that merging occupied a place somewhere in between combinatorial voting [Lang and Xia, 2016] and multiwinner voting [Faliszewski et al., 2017a], and that the transfer of knowledge from the classical voting models to more sophisticated settings was a matter of considerable interest, so we set our sights on strategyproofness. At the same time, our interests were equally stoked by the idea that merging, or a merging-like framework, could be used to aggregate other types of formalisms of interest to the AI community, such as Horn formulas [Haret et al., 2015, Haret et al., 2017] or abstract argumentation frameworks [Delobelle et al., 2016]. This led us to consider applying acceptance notions (such as the skeptical and credulous notions presented in Sections 5.4) to the results of a merging operator, and to see what

happened to the existing strategyproofness results [Everaere et al., 2007]. Since our methods for calculating satisfaction with respect to the merging results were different from the original setting [Everaere et al., 2007], there was no promise that its results would be instantly applicable. What we found, however, was that the situation was even worse, in the sense that, with one exception, restrictions that guaranteed strategyproofness in [Everaere et al., 2007] failed to do so in our setting.

The main goal for future research here is to tie the properties in Sections 5.1, 5.2 and 5.3 together with the notions of strategyproofness in Section 5.4 for a general result along the lines of the classical theorems of social choice theory [Gibbard, 1973, Satterthwaite, 1975, Duggan and Schwartz, 2000]. Our aim is to also consider extended settings of manipulation, e.g., bribery [Baumeister et al., 2015], where sets of agents can be incentivized to form a joint manipulating coalition. Our work on merging and proportionality also suggests several directions for future research. Even though the two proportionality postulates $\mathrm{M}_{\mathsf{CPROP}}$ and $\mathrm{M}_{\mathsf{BPROP}}$ we proposed apply only to very restricted instances, experience has shown that even weak proportionality postulates have proven sufficient for axiomatic characterizations [Lackner and Skowron, 2018b]. In our work, as well, these two postulates are sufficient to distinguish proportional from non-proportional operators. On the other hand, stronger postulates are desirable to determine to which degree proportionality guarantees can be given. This has recently been investigated in the context of approval-based committee elections [Aziz et al., 2017, Aziz et al., 2018, Fernández et al., 2017], and this line of work can serve as a basis for a similar analysis for belief merging operators. Coming back to manipulation, it can be fully expected that proportional belief merging operators are prone to strategic voting, as in the setting of approval-based committee elections even weak forms of proportionality and strategy-proofness have been shown to be incompatible [Peters, 2018]. Still, it has been found that the percentage of manipulable instances depends strongly on the choice of voting rules [Lackner and Skowron, 2018a], indicating that a detailed analysis of vulnerabilities is an interesting avenue for future work. Finally, it would be interesting to see if the framework of merging can be used in different social choice contexts, e.g., resource allocation [Chevaleyre et al., 2017].

Chapters 4 and 5 are both concerned with foundational issues in the theory of belief change. The remaining chapters have a more applied bent. Chapter 6 takes us back to the single-agent belief change operations of revision and update, this time applied to the Horn fragment. Section 6.2 developed alongside Chapter 4. It was clear to us that in certain situations postulate R_2 would make an HPH-revision operator choose a set of interpretations that could not be expressed as a Horn formula, but a weaker version of postulate R_2, which allowed the operator to select some portion of that set that could be expressed as a Horn formula, might work. The catch, of course, was that such a discriminatory behavior was bound to violate postulate $\mathrm{R}_{\mathsf{NEUT}}$. Sections 6.3 and 6.4 grew out of an attempt to extend the basic framework for revision in fragments in [Delgrande and Peppas, 2015, Delgrande et al., 2018] to other settings: first of all to update, and then to the weaker postulates R_{7-8} (and U_{7-8}), describing partial preorders. The latter

turned out to be more challenging. The main challenge for the future, in this case, is to flesh out the properties that are essential for the representation results to work, and extend these results to other fragments, in the manner of existing models [Delgrande et al., 2018].

Chapter 7 applies the principles of belief change to preferences. Since belief change, as described in Chapter 3, is itself about choice and preferences, preference change ended up being characterized in terms of preferences on preferences, which coincided with thoughts about the dynamics of preferences from Sen and others [Sen, 1977]. Interestingly, the principles that were best suited for this type of operation were not the revision postulates R_{1-6}, but their dual versions E_{1-6}, used for enforcement. This formulation also suggested the right kind of choice function for preference revision operators, with the addition operator in Chapter 7 being adapted directly from the addition operator in Section 3.3. In general, due to its flipped choice function that adds elements to the new information rather than removing them, enforcement is more suited to describe the dynamics of types of objects that are constructed out of some building block-like elements, in the way in which strict partial orders are constructed out of their comparisons. By contrast, a propositional formula is not 'made up of' its models in the same way in which a partial order is constructed out of its comparisons: a propositional formula is more like a set of specifications, with its models being the outcomes that meet those specifications.

To put this differently, we could have have approached preference revision in an alternative way, by using a logical formalism in which the object being revised would be something like a preference formula φ, whose models are all the different preference orders that satisfy it. Such a formalism could be a fragment of propositional logic, e.g., of acyclic definite Horn clauses consisting of two variables, where one such clause $a \to b$ encodes the comparison that a is at least as good as b. Such a fragment, however, is not closed under conjunction, and therefore does not fall within the purview of existing work on revision in fragments [Delgrande et al., 2018]. Another option would be to use a formalism specifically tailored to talk about preference orders, such as the language PL [Bienvenu et al., 2010], but there we would encounter the same problems of expressibility, i.e., of making sure that the output will be expressible in the target language. Either way, if we proceeded in this way, the revision problem would have amounted to selecting some models of the new information μ, and in this case the revision postulate R_{1-6} would be the appropriate postulates to use. This is definitely a viable alternative, and a promising direction for further work.

More generally, one lesson that can be drawn from this thesis is that a belief change operator arises out of a combination of a few basic elements: a language for representing the information, a set of logical postulates, a set of semantic properties describing the preferences over outcomes, and a choice procedure connecting the two. For propositional revision we have propositional logic, postulates R_{1-8}, properties r_{1-7} and the choice procedure that selects the minimal elements of μ. In the Horn fragment we have the same choice procedures but a different representation language, which then requires that the postulates and properties be supplemented to make up for the expressive limitations of

Horn formulas. For enforcement and its offshoots, the postulates are E_{1-6}, the properties are e_{1-6} and the choice function is given by the addition operator, with additional quirks depending on the type of representation language used. Designing a belief change operator requires all these elements to work together, in what is usually a delicate and fragile balance: modifying one element, even slightly, usually requires rethinking most of the other elements as well. At the moment, a universal, foolproof recipe for applying belief change to any Knowledge Representation formalism we might be interested in still seems slightly out of reach. Hopefully, as more work becomes available, the gap will become narrower.

Bibliography

[3blue1brown, 2019] 3blue1brown (2019). Bayes' theorem, and making probability intuitive. https://www.youtube.com/watch?v=HZGCoVF3YvM.

[Abelson, 1986] Abelson, R. (1986). Beliefs Are Like Possessions. *Journal for the Theory of Social Behaviour*, 16:223 – 250.

[Alchourrón et al., 1985] Alchourrón, C. E., Gärdenfors, P., and Makinson, D. (1985). On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *The Journal of Symbolic Logic*, 50(2):510–530.

[Alchourrón and Makinson, 1985] Alchourrón, C. E. and Makinson, D. (1985). On the logic of theory change: Safe contraction. *Studia Logica*, 44(4):405–422.

[Anand, 2009] Anand, P. (2009). Rationality and Intransitive Preference: Foundations for the Modern View. In Anand, P., Pattanaik, P., and Puppe, C., editors, *Handbook of Rational and Social Choice*. Oxford University Press.

[Arló-Costa and Pedersen, 2010] Arló-Costa, H. and Pedersen, A. P. (2010). Social Norms, Rational Choice and Belief Change. In Olsson, E. J. and Enqvist, S., editors, *Belief Revision meets Philosophy of Science*, pages 163–212. Springer.

[Arrow, 1951] Arrow, K. J. (1951). *Social Choice and Individual Values*. Yale University Press, First edition.

[Arrow, 1959] Arrow, K. J. (1959). Rational Choice Functions and Orderings. *Economica*, 26(102):121–127.

[Aziz et al., 2017] Aziz, H., Brill, M., Conitzer, V., Elkind, E., Freeman, R., and Walsh, T. (2017). Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2):461–485.

[Aziz et al., 2018] Aziz, H., Elkind, E., Huang, S., Lackner, M., Fernández, L. S., and Skowron, P. (2018). On the Complexity of Extended and Proportional Justified Representation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 902–909.

[Balinski and Young, 1982] Balinski, M. and Young, H. P. (1982). *Fair Representation: Meeting the Ideal of One Man, One Vote*. Yale University Press.

[Baltag et al., 2019] Baltag, A., Gierasimczuk, N., and Smets, S. (2019). Truth-Tracking by Belief Revision. *Studia Logica*, 107(5):917–947.

[Bar-Hillel and Margalit, 1988] Bar-Hillel, M. and Margalit, A. (1988). How Vicious Are Cycles of Intransitive Choice? *Theory and Decision*, 24(2):119–145.

[Baumann, 2012] Baumann, R. (2012). What Does it Take to Enforce an Argument? Minimal Change in Abstract Argumentation. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012)*, pages 127–132.

[Baumeister et al., 2015] Baumeister, D., Erdélyi, G., Erdélyi, O. J., and Rothe, J. (2015). Complexity of manipulation and bribery in judgment aggregation for uniform premise-based quota rules. *Mathematical Social Sciences*, 76:19–30.

[Baumeister and Rothe, 2016] Baumeister, D. and Rothe, J. (2016). Preference Aggregation by Voting. In Rothe, J., editor, *Economics and Computation: An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division*, pages 197–325. Springer.

[Baumeister et al., 2017] Baumeister, D., Rothe, J., and Selker, A.-K. (2017). Strategic behavior in judgment aggregation. In Endriss, U., editor, *Trends in Computational Social Choice*, pages 145–168. AI Access.

[Benferhat et al., 2005] Benferhat, S., Lagrue, S., and Papini, O. (2005). Revision of Partially Ordered Information: Axiomatization, Semantics and Iteration. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 376–381.

[Benthem and Liu, 2014] Benthem, J. and Liu, F. (2014). Deontic Logic and Preference Change. *IfCoLog Journal of Logics and their Applications*, 1(2):1–46.

[Bienvenu et al., 2010] Bienvenu, M., Lang, J., and Wilson, N. (2010). From Preference Logics to Preference Languages, and Back. In *Proceedings of the Twelfth International Conference on Principles of Knowledge Representation and Reasoning (KR 2010)*, pages 414–424.

[Binnewies et al., 2018] Binnewies, S., Zhuang, Z., Wang, K., and Stantic, B. (2018). Syntax-Preserving Belief Change Operators for Logic Programs. *ACM Transactions of Computational Logic*, 19(2):12:1–12:42.

[Black, 1958] Black, D. (1958). *The Theory of Committees and Elections*. Cambridge University Press.

[Bonanno, 2009] Bonanno, G. (2009). Rational choice and AGM belief revision. *Artificial Intelligence*, 173(12-13):1194–1203.

[Booth et al., 2009] Booth, R., Meyer, T. A., and Varzinczak, I. J. (2009). Next Steps in Propositional Horn Contraction. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 702–707.

[Booth et al., 2011] Booth, R., Meyer, T. A., Varzinczak, I. J., and Wassermann, R. (2011). On the Link between Partial Meet, Kernel, and Infra Contraction and its Application to Horn Logic. *Journal of Artificial Intelligence Research (JAIR)*, 42:31–53.

[Bossert and Suzumura, 2010] Bossert, W. and Suzumura, K. (2010). *Consistency, Choice, and Rationality*. Harvard University Press.

[Boutilier et al., 2004] Boutilier, C., Brafman, R. I., Domshlak, C., Hoos, H. H., and Poole, D. (2004). CP-nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements. *Journal of Artificial Intelligence Research (JAIR)*, 21:135–191.

[Bradley, 2007] Bradley, R. (2007). The kinematics of belief and desire. *Synthese*, 156(3):513–535.

[Brams et al., 2007] Brams, S. J., Kilgour, M. D., and Sanver, R. M. (2007). A minimax procedure for electing committees. *Public Choice*, 132(3):401–420.

[Brandt et al., 2016] Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors (2016). *Handbook of Computational Social Choice*. Cambridge University Press.

[Brody, 2020] Brody, R. (2020). The Fossilized 2020 Oscar Nominations. Online at `https://www.newyorker.com/culture/the-front-row/the-fossilized-2020-oscar-nominations`.

[Cadilhac et al., 2015] Cadilhac, A., Asher, N., Lascarides, A., and Benamara, F. (2015). Preference change. *Journal of Logic, Language and Information*, 24(3):267–288.

[Caridroit et al., 2017] Caridroit, T., Konieczny, S., and Marquis, P. (2017). Contraction in propositional logic. *International Journal of Approximate Reasoning*, 80:428–442.

[Chernoff, 1954] Chernoff, H. (1954). Rational selection of decision functions. *Econometrica*, pages 422–443.

[Chevaleyre et al., 2008] Chevaleyre, Y., Endriss, U., Lang, J., and Maudet, N. (2008). Preference Handling in Combinatorial Domains: From AI to Social Choice. *AI Magazine*, 29(4):37–46.

[Chevaleyre et al., 2017] Chevaleyre, Y., Endriss, U., and Maudet, N. (2017). Distributed fair allocation of indivisible goods. *Artificial Intelligence*, 242:1–22.

[Chomicki and Song, 2005] Chomicki, J. and Song, J. (2005). Monotonic and Nonmonotonic Preference Revision. In *Proc. IJCAI 2005 Multidisciplinary Workshop on Advances in Preference Handling*.

[Chopra et al., 2006] Chopra, S., Ghose, A. K., and Meyer, T. A. (2006). Social choice theory, belief merging, and strategy-proofness. *Information Fusion*, 7(1):61–79.

[Conitzer and Walsh, 2016] Conitzer, V. and Walsh, T. (2016). Barriers to Manipulation in Voting. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors, *Handbook of Computational Social Choice*, pages 127–145. Cambridge University Press.

[Creignou et al., 2018a] Creignou, N., Haret, A., Papini, O., and Woltran, S. (2018a). Belief Update in the Horn Fragment. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)*, pages 1781–1787.

[Creignou et al., 2018b] Creignou, N., Ktari, R., and Papini, O. (2018b). Belief Update within Propositional Fragments. *Journal of Artificial Intelligence Research (JAIR)*, 61:807–834.

[Creignou et al., 2014] Creignou, N., Papini, O., Pichler, R., and Woltran, S. (2014). Belief revision within fragments of propositional logic. *Journal of Computer and System Sciences*, 80(2):427–449.

[Creignou et al., 2016] Creignou, N., Papini, O., Rümmele, S., and Woltran, S. (2016). Belief Merging within Fragments of Propositional Logic. *ACM Transactions of Computational Logic*, 17(3):20:1–20:28.

[Dalal, 1988] Dalal, M. (1988). Investigations into a Theory of Knowledge Base Revision. In *Proceedings of the 7th National Conference on Artificial Intelligence, 1988*, pages 475–479.

[Darwiche and Pearl, 1997] Darwiche, A. and Pearl, J. (1997). On the Logic of Iterated Belief Revision. *Artificial Intelligence*, 89(1-2):1–29.

[Delgrande, 2008] Delgrande, J. P. (2008). Horn Clause Belief Change: Contraction Functions. In *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR 2008)*, pages 156–165.

[Delgrande and Peppas, 2015] Delgrande, J. P. and Peppas, P. (2015). Belief revision in Horn theories. *Artificial Intelligence*, 218:1–22.

[Delgrande et al., 2018] Delgrande, J. P., Peppas, P., and Woltran, S. (2018). General Belief Revision. *Journal of the ACM (JACM)*, 65(5):29:1–29:34.

[Delgrande and Wassermann, 2010] Delgrande, J. P. and Wassermann, R. (2010). Horn Clause Contraction Functions: Belief Set and Belief Base Approaches. In *Proceedings of the Twelfth International Conference on Principles of Knowledge Representation and Reasoning (KR 2010)*, pages 143–152.

[Delgrande and Wassermann, 2013] Delgrande, J. P. and Wassermann, R. (2013). Horn Clause Contraction Functions. *Journal of Artificial Intelligence Research (JAIR)*, 48:475–511.

[Delobelle et al., 2016] Delobelle, J., Haret, A., Konieczny, S., Mailly, J., Rossit, J., and Woltran, S. (2016). Merging of Abstract Argumentation Frameworks. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2016)*, pages 33–42.

[Deza and Deza, 2016] Deza, M. M. and Deza, E. (2016). *Encyclopedia of Distances*. Springer, 4 edition.

[Díaz and Pino Pérez, 2017] Díaz, A. M. and Pino Pérez, R. (2017). Impossibility in belief merging. *Artificial Intelligence*, 251:1–34.

[Díaz and Pino Pérez, 2018] Díaz, A. M. and Pino Pérez, R. (2018). Epistemic states, fusion and strategy-proofness. In *Proceedings of the 17th International Workshop on Non-Monotonic Reasoning (NMR 2018)*, pages 176–185.

[Diller et al., 2015] Diller, M., Haret, A., Linsbichler, T., Rümmele, S., and Woltran, S. (2015). An Extension-Based Approach to Belief Revision in Abstract Argumentation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 2926–2932.

[Diller et al., 2018] Diller, M., Haret, A., Linsbichler, T., Rümmele, S., and Woltran, S. (2018). An extension-based approach to belief revision in abstract argumentation. *International Journal of Approximate Reasoning*, 93:395–423.

[Domshlak et al., 2011] Domshlak, C., Hüllermeier, E., Kaci, S., and Prade, H. (2011). Preferences in AI: An Overview. *Artificial Intelligence*, 175(7-8):1037–1052.

[Doyle, 1991] Doyle, J. (1991). Rational Belief Revision (Preliminary Report). In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR 1991)*, pages 163–174.

[Dubois et al., 2017] Dubois, D., Lorini, E., and Prade, H. (2017). The Strength of Desires: A Logical Approach. *Minds and Machines*, 27(1):199–231.

[Duggan and Schwartz, 2000] Duggan, J. and Schwartz, T. (2000). Strategic manipulability without resoluteness or shared beliefs: Gibbard-Satterthwaite generalized. *Social Choice and Welfare*, 17(1):85–93.

[Dummett, 1984] Dummett, M. (1984). *Voting Procedures*. Oxford University Press.

[Eckert and Pigozzi, 2005] Eckert, D. and Pigozzi, G. (2005). Belief merging, judgment aggregation and some links with social choice theory. In *Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics. Dagstuhl Seminar Proceedings*.

[Eiter and Gottlob, 1992] Eiter, T. and Gottlob, G. (1992). On the Complexity of Propositional Knowledge Base Revision, Updates, and Counterfactuals. *Artificial Intelligence*, 57(2-3):227–270.

[Endriss, 2011] Endriss, U. (2011). Applications of logic in social choice theory - (invited talk). In *Proceedings of the 12th Workshop on Computational Logic in Multi-Agent Systems (CLIMA 2011)*, pages 88–91.

[Endriss, 2016] Endriss, U. (2016). Judgment Aggregation. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors, *Handbook of Computational Social Choice*, pages 399–426. Cambridge University Press.

[Endriss, 2017] Endriss, U., editor (2017). *Trends in Computational Social Choice*. AI Access.

[Everaere et al., 2007] Everaere, P., Konieczny, S., and Marquis, P. (2007). The Strategy-Proofness Landscape of Merging. *Journal of Artificial Intelligence Research (JAIR)*, 28:49–105.

[Everaere et al., 2010a] Everaere, P., Konieczny, S., and Marquis, P. (2010a). Disjunctive merging: Quota and Gmin merging operators. *Artificial Intelligence*, 174(12-13):824–849.

[Everaere et al., 2010b] Everaere, P., Konieczny, S., and Marquis, P. (2010b). The Epistemic View of Belief Merging: Can We Track the Truth? In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010)*, pages 621–626.

[Everaere et al., 2014] Everaere, P., Konieczny, S., and Marquis, P. (2014). On Egalitarian Belief Merging. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2014)*, pages 121–140.

[Everaere et al., 2017] Everaere, P., Konieczny, S., and Marquis, P. (2017). An Introduction to Belief Mergingand its Links with JudgmentAggregation. In Endriss, U., editor, *Trends in Computational Social Choice*. AI Access.

[Faliszewski et al., 2017a] Faliszewski, P., Skowron, P., Slinko, A., and Talmon, N. (2017a). Multiwinner Voting: A New Challenge for Social Choice Theory. In Endriss, U., editor, *Trends in Computational Social Choice*, pages 27–47. AI Access Foundation.

[Faliszewski et al., 2017b] Faliszewski, P., Slinko, A., and Talmon, N. (2017b). The Complexity of Multiwinner Voting Rules with Variable Number of Winners. *CoRR*, abs/1711.06641.

[Fermé and Hansson, 2018] Fermé, E. L. and Hansson, S. O. (2018). *Belief Change: Introduction and Overview*. Springer Briefs in Intelligent Systems. Springer.

[Fernández et al., 2017] Fernández, L. S., Elkind, E., Lackner, M., García, N. F., Arias-Fisteus, J., Basanta-Val, P., and Skowron, P. (2017). Proportional Justified Representation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, pages 670–676.

[Foley, 1993] Foley, R. (1993). *Working Without a Net: A Study of Egocentric Epistemology*. Oxford University Press.

[Forbus, 1989] Forbus, K. D. (1989). Introducing Actions into Qualitative Simulation. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI) 1989*, pages 1273–1278.

[Frankfurt, 1988] Frankfurt, H. G. (1988). Freedom of the Will and the Concept of a Person. In *What is a person?*, pages 127–144. Springer.

[Freund, 2004] Freund, M. (2004). On the revision of preferences and rational inference processes. *Artificial Intelligence*, 152(1):105–137.

[Gabbay et al., 2007] Gabbay, D. M., Pigozzi, G., and Rodrigues, O. (2007). Common foundations for belief revision, belief merging and voting. In *Formal Models of Belief Change in Rational Agents. Dagstuhl Seminar Proceedings*.

[Gärdenfors, 1988] Gärdenfors, P. (1988). *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. The MIT Press.

[Gärdenfors, 2011] Gärdenfors, P. (2011). Notes on the History of Ideas Behind AGM. *Journal of Philosophical Logic*, 40(2):115–120.

[Gärdenfors and Makinson, 1988] Gärdenfors, P. and Makinson, D. (1988). Revisions of Knowledge Systems Using Epistemic Entrenchment. In *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge (TARK 1988)*, pages 83–95.

[Gibbard, 1973] Gibbard, A. (1973). Manipulation of Voting Schemes: A General Result. *Econometrica*, pages 587–601.

[Goodall, 2010] Goodall, J. (2010). *Through a Window: My Thirty Years With the Chimpanzees of Gombe*. Mariner Books: Houghton Mifflin Harcourt.

[Grant and Zandt, 2009] Grant, S. and Zandt, T. V. (2009). Expected Utility Theory. In Anand, P., Pattanaik, P., and Puppe, C., editors, *Handbook of Rational and Social Choice*. Oxford University Press.

[Grove, 1988] Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic*, 17(2):157–170.

[Grüne-Yanoff, 2013] Grüne-Yanoff, T. (2013). Preference change and conservatism: comparing the bayesian and the AGM models of preference revision. *Synthese*, 190(14):2623–2641.

[Grüne-Yanoff and Hansson, 2009] Grüne-Yanoff, T. and Hansson, S. O. (2009). From Belief Revision to Preference Change. In *Preference Change: Approaches from Philosophy, Economics and Psychology*, pages 159–184. Springer.

[Grüne-Yanoff and Hansson, 2009] Grüne-Yanoff, T. and Hansson, S. O., editors (2009). *Preference Change: Approaches from Philosophy, Economics and Psychology*, volume 42 of *Theory and Decision Library A*. Springer.

[Hansson, 1968] Hansson, B. (1968). Choice Structures and Preference Relations. *Synthese*, 18(4):443–458.

[Hansson, 1995] Hansson, S. O. (1995). Changes in preference. *Theory and Decision*, 38(1):1–28.

[Hansson, 1999a] Hansson, S. O. (1999a). A Survey of non-Prioritized Belief Revision. *Erkenntnis*, 50(2-3):413–427.

[Hansson, 1999b] Hansson, S. O. (1999b). *A Textbook of Belief Dynamics: Theory Change and Database Updating*, volume 11 of *Applied logic series*. Kluwer.

[Hansson, 2014] Hansson, S. O. (2014). Descriptor Revision. *Studia Logica*, 102(5):955–980.

[Hansson, 2017] Hansson, S. O. (2017). Logic of Belief Revision. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition.

[Hansson et al., 2001] Hansson, S. O., Fermé, E. L., Cantwell, J., and Falappa, M. A. (2001). Credibility Limited Revision. *The Journal of Symbolic Logic*, 66(4):1581–1596.

[Hansson and Grüne-Yanoff, 2018] Hansson, S. O. and Grüne-Yanoff, T. (2018). Preferences. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition.

[Haret et al., 2018a] Haret, A., Khani, H., Moretti, S., and Öztürk, M. (2018a). Ceteris paribus majority for social ranking. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)*, pages 303–309.

[Haret et al., 2020] Haret, A., Lackner, M., Pfandler, A., and Wallner, J. P. (2020). Proportional Belief Merging. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*.

[Haret et al., 2016a] Haret, A., Mailly, J., and Woltran, S. (2016a). Distributing Knowledge into Simple Bases. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 2016)*, pages 1109–1115.

[Haret et al., 2018b] Haret, A., Novaro, A., and Grandi, U. (2018b). Preference Aggregation with Incomplete CP-Nets. In *Proceedings of the Sixteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2018)*, pages 308–318.

[Haret et al., 2016b] Haret, A., Pfandler, A., and Woltran, S. (2016b). Beyond IC Postulates: Classification Criteria for Merging Operators. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI 2016)*, pages 372–380.

[Haret et al., 2015] Haret, A., Rümmele, S., and Woltran, S. (2015). Merging in the horn fragment. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 3041–3047.

[Haret et al., 2017] Haret, A., Rümmele, S., and Woltran, S. (2017). Merging in the Horn fragment. *ACM Transactions of Computational Logic*, 18(1):6:1–6:32.

[Haret and Wallner, 2019] Haret, A. and Wallner, J. P. (2019). Manipulating Skeptical and Credulous Consequences When Merging Beliefs. In *Proceedings of the 16th European Conference on Logics in Artificial Intelligence (JELIA 2019)*, pages 133–150.

[Haret et al., 2018c] Haret, A., Wallner, J. P., and Woltran, S. (2018c). Two Sides of the Same Coin: Belief Revision and Enforcing Arguments. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)*, pages 1854–1860.

[Haret and Woltran, 2017] Haret, A. and Woltran, S. (2017). Deviation in Belief Change on Fragments of Propositional Logic. In *Proceedings of the 6th Workshop on Dynamics of Knowledge and Belief (DKB-2017) and the 5th Workshop KI & Kognition (KIK-2017) co-located with 40th German Conference on Artificial Intelligence (KI 2017)*, pages 64–76.

[Haret and Woltran, 2018] Haret, A. and Woltran, S. (2018). Belief Revision Operators with Varying Attitudes Towards Initial Beliefs. In *Proceedings of the 17th International Workshop on Non-Monotonic Reasoning (NMR 2018)*, pages 156–165.

[Haret and Woltran, 2019] Haret, A. and Woltran, S. (2019). Belief Revision Operators with Varying Attitudes Towards Initial Beliefs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pages 1726–1733.

[Harper, 1976] Harper, W. L. (1976). Rational Conceptual Change. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1976:462–494.

[Harsanyi, 1955] Harsanyi, J. C. (1955). Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy*, 63(4):309–321.

[Hausman, 2011] Hausman, D. M. (2011). *Preference, Value, Choice, and Welfare*. Cambridge University Press.

[Herzberger, 1973] Herzberger, H. G. (1973). Ordinal Preference and Rational Choice. *Econometrica*, 41(2):187–237.

[Herzig and Rifi, 1999] Herzig, A. and Rifi, O. (1999). Propositional Belief Base Update and Minimal Change. *Artificial Intelligence*, 115(1):107–138.

[Horn, 1951] Horn, A. (1951). On Sentences Which are True of Direct Unions of Algebras. *The Journal of Symbolic Logic*, 16(1):14–21.

[Jackson, 2020] Jackson, E. G. (2020). The relationship between belief and credence. *Philosophy Compass*.

[Jeffrey, 1974] Jeffrey, R. C. (1974). Preference among preferences. *Journal of Philosophy*, 71(13):377–391.

[Joyce, 2019] Joyce, J. (2019). Bayes' Theorem. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2019 edition.

[Kahneman, 2011] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[Katsuno and Mendelzon, 1991] Katsuno, H. and Mendelzon, A. O. (1991). On the Difference between Updating a Knowledge Base and Revising It. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 387–394.

[Katsuno and Mendelzon, 1992] Katsuno, H. and Mendelzon, A. O. (1992). Propositional Knowledge Base Revision and Minimal Change. *Artificial Intelligence*, 52(3):263–294.

[Kelly, 1998] Kelly, K. T. (1998). The Learning Power of Belief Revision. In *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-98)*, pages 111–124.

[Kilgour, 2016] Kilgour, M. D. (2016). Approval elections with a variable number of winners. *Theory and Decision*, 81(2):199–211.

[Konieczny et al., 2004] Konieczny, S., Lang, J., and Marquis, P. (2004). DA ^{2} merging operators. *Artificial Intelligence*, 157(1-2):49–79.

[Konieczny and Pino Pérez, 2002] Konieczny, S. and Pino Pérez, R. (2002). Merging Information Under Constraints: A Logical Framework. *Journal of Logic and Computation*, 12(5):773–808.

[Konieczny and Pino Pérez, 2005] Konieczny, S. and Pino Pérez, R. (2005). Propositional belief base merging or how to merge beliefs/goals coming from several sources and some links with social choice theory. *European Journal of Operational Research*, 160(3):785–802.

[Konieczny and Pino Pérez, 2011] Konieczny, S. and Pino Pérez, R. (2011). Logic Based Merging. *Journal of Philosophical Logic*, 40(2):239–270.

[Kraus et al., 1990] Kraus, S., Lehmann, D., and Magidor, M. (1990). Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *Artif. Intell.*, 44(1-2):167–207.

[Lackner and Skowron, 2018a] Lackner, M. and Skowron, P. (2018a). Approval-Based Multi-Winner Rules and Strategic Voting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)*, pages 340–346.

[Lackner and Skowron, 2018b] Lackner, M. and Skowron, P. (2018b). Consistent Approval-Based Multi-Winner Rules. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC 2018)*, pages 47–48.

[Lang and van der Torre, 2008] Lang, J. and van der Torre, L. W. N. (2008). Preference Change Triggered by Belief Change: A Principled Approach. In *Proceedings of the 8th International Conference on Logic and the Foundations of Game and Decision Theory (LOFT 8)*, pages 86–111.

[Lang and Xia, 2016] Lang, J. and Xia, L. (2016). Voting in Combinatorial Domains. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors, *Handbook of Computational Social Choice*, pages 197–222. Cambridge University Press.

[Levesque, 1986] Levesque, H. J. (1986). Making Believers out of Computers. *Artificial Intelligence*, 30(1):81–108.

[Levi, 1980] Levi, I. (1980). *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. The MIT press.

[Levi, 1991] Levi, I. (1991). *The Fixation of Belief and Its Undoing: Changing Beliefs Through Inquiry*. Cambridge University Press.

[Liberatore and Schaerf, 1998] Liberatore, P. and Schaerf, M. (1998). Arbitration (or How to Merge Knowledge Bases). *IEEE Transactions on Knowledge and Data Engineering*, 10(1):76–90.

[Liu, 2011] Liu, F. (2011). *Reasoning About Preference Dynamics*, volume 354 of *Synthese Library*. Springer.

[Luce, 1956] Luce, R. D. (1956). Semiorders and a Theory of Utility Discrimination. *Econometrica*, 24(2):178–191.

[Luce and Raiffa, 1957] Luce, R. D. and Raiffa, H. (1957). *Games and Decisions: Introduction and Critical Survey*. John Wiley & Sons, Inc.

[Ma et al., 2012] Ma, J., Benferhat, S., and Liu, W. (2012). Revising Partial Pre-Orders with Partial Pre-Orders: A Unit-Based Revision Framework. In *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning (KR 2012)*, pages 633–637.

[Ma et al., 2015] Ma, J., Liu, W., and Dubois, D. (2015). Rational Partial Choice Functions and Their Application to Belief Revision. In *Proceedings of the 8th International Knowledge Science, Engineering and Management Conference, KSEM 2015*, pages 128–140.

[Marquis and Schwind, 2014] Marquis, P. and Schwind, N. (2014). Lost in translation: Language independence in propositional logic-application to belief change. *Artificial Intelligence*, 206:1–24.

[Mas-Colell et al., 1995] Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic Theory*. Oxford University Press.

[Maynard-Zhang and Lehmann, 2003] Maynard-Zhang, P. and Lehmann, D. (2003). Representing and Aggregating Conflicting Beliefs. *Journal of Artificial Intelligence Research (JAIR)*, 19:155–203.

[McKinsey, 1943] McKinsey, J. C. C. (1943). The Decision Problem for Some Classes of Sentences Without Quantifiers. *The Journal of Symbolic Logic*, 8(2):61–76.

[Meyer, 2001] Meyer, T. A. (2001). On the semantics of combination operations. *Journal of Applied Non-Classical Logics*, 11(1-2):59–84.

[Meyer et al., 2001] Meyer, T. A., Ghose, A., and Chopra, S. (2001). Social Choice, Merging, and Elections. In *Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2001)*, pages 466–477.

[Monroe, 1995] Monroe, B. L. (1995). Fully Proportional Representation. *American Political Science Review*, 89(4):925–940.

[Nash, 1950] Nash, J. F. (1950). The Bargaining Problem. *Econometrica*, pages 155–162.

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

[Nozick, 1994] Nozick, R. (1994). *The Nature of Rationality*. Princeton University Press.

[Olsson, 2003] Olsson, E. J. (2003). Belief Revision, Rational Choice and the Unity of Reason. *Studia Logica*, 73(2):219–240.

[Pagnucco et al., 1994] Pagnucco, M., Nayak, A. C., and Foo, N. Y. (1994). Abductive Expansion: Abductive Inference and the Process of Belief Change. In *Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence (AI94)*, pages 267–274.

[Pearl, 1989] Pearl, J. (1989). Probabilistic Semantics for Nonmonotonic Reasoning: A Survey. In *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning (KR'89)*, pages 505–516.

[Peppas, 2008] Peppas, P. (2008). Belief Revision. In van Harmelen, F., Lifschitz, V., and Porter, B. W., editors, *Handbook of Knowledge Representation*, pages 317–359. Elsevier.

[Peppas et al., 1996] Peppas, P., Nayak, A. C., Pagnucco, M., Foo, N. Y., Kwok, R. B. H., and Prokopenko, M. (1996). Revision vs. Update: Taking a Closer Look. In Wahlster, W., editor, *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI 1996)*, pages 95–99. John Wiley and Sons, Chichester.

[Peppas and Williams, 2016] Peppas, P. and Williams, M. (2016). Kinetic Consistency and Relevance in Belief Revision. In *Proceedings of the 15th European Conference on Logics in Artificial Intelligence (JELIA 2016)*, pages 401–414.

[Peters, 2018] Peters, D. (2018). Proportionality and Strategyproofness in Multiwinner Elections. In André, E., Koenig, S., Dastani, M., and Sukthankar, G., editors, *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, pages 1549–1557.

[Pfandler et al., 2015] Pfandler, A., Rümmele, S., Wallner, J. P., and Woltran, S. (2015). On the Parameterized Complexity of Belief Revision. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 3149–3155.

[Pigozzi et al., 2016] Pigozzi, G., Tsoukiàs, A., and Viappiani, P. (2016). Preferences in artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, 77(3-4):361–401.

[Quinn, 1990] Quinn, W. S. (1990). The Puzzle of the Self-Torturer. *Philosophical Studies*, 59(1):79–90.

[Radner and Marschak, 1954] Radner, R. and Marschak, J. (1954). Note on Some Proposed Decision Criteria. In Thrall, R. M., Coombs, C. H., and Davis, R. L., editors, *Decision Processes*, pages 61–68. Wiley.

[Rossi and Mattei, 2019] Rossi, F. and Mattei, N. (2019). Building Ethically Bounded AI. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*, pages 9785–9789.

[Rossi et al., 2011] Rossi, F., Venable, K. B., and Walsh, T. (2011). *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

[Rott, 1992] Rott, H. (1992). Modellings for Belief Change: Base Contraction, Multiple Contraction, and Epistemic Entrenchment. In *Proceedings of the European Workshop on Logics in AI (JELIA '92)*, pages 139–153.

[Rott, 1993] Rott, H. (1993). Belief Contraction in the Context for the General Theory of Rational Choice. *The Journal of Symbolic Logic*, 58(4):1426–1450.

[Rott, 2001] Rott, H. (2001). *Change, Choice and Inference: A study of Belief Revision and Nonmonotonic Reasoning*, volume 42 of *Oxford Logic Guides*. Oxford University Press.

[Ryan, 1996] Ryan, M. (1996). Belief Revision and Ordered Theory Presentations. In *Logic, Action, and Information*, pages 129–151.

[Satterthwaite, 1975] Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217.

[Schulte, 1999] Schulte, O. (1999). Minimal Belief Change and the Pareto Principle. *Synthese*, 118(3):329–361.

[Schwind et al., 2018] Schwind, N., Konieczny, S., and Marquis, P. (2018). On Belief Promotion. In *Proceedings of the Sixteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2018)*, pages 297–307.

[Schwitzgebel, 2019] Schwitzgebel, E. (2019). Belief. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition.

[Sen, 1969] Sen, A. K. (1969). Quasi-Transitivity, Rational Choice and Collective Decisions. *The Review of Economic Studies*, 36(3):381–393.

[Sen, 1970] Sen, A. K. (1970). *Collective Choice and Social Welfare*. Holden-Day.

[Sen, 1977] Sen, A. K. (1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs*, pages 317–344.

[Sen, 2017] Sen, A. K. (2017). *Collective Choice and Social Welfare: Expanded Edition*. Penguin UK.

[Shoham, 1987] Shoham, Y. (1987). *A Semantical Approach to Nonmonotonic Logics*, page 227–250. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Strasser and Antonelli, 2019] Strasser, C. and Antonelli, A. G. (2019). Non-monotonic Logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition.

[Suzumura, 1976] Suzumura, K. (1976). Remarks on the Theory of Collective Choice. *Economica*, 43(172):381–390.

[Suzumura, 1983] Suzumura, K. (1983). *Rational Choice, Collective Decisions and Social Welfare*. Cambridge University Press.

[Suzumura, 2016] Suzumura, K. (2016). *Choice, Preferences, and Procedures: A Rational Choice Theoretic Approach*. Harvard University Press.

[Szpilrajn, 1930] Szpilrajn, E. (1930). Sur l'extension de l'ordre partiel. *Fundamenta Mathematicae*, 16(1):386–389.

[Thiele, 1895] Thiele, T. N. (1895). Om flerfoldsvalg. In *Oversigt over det Kongelige Danske Videnskabernes Selskabs Forhandlinger*, pages 415–441.

[Tideman, 2006] Tideman, N. (2006). *Collective Decisions and Voting: The Potential for Public Choice*. Ashgate Publishing, Ltd.

[Tversky and Kahneman, 1981] Tversky, A. and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481):453–458.

[Wallner et al., 2017] Wallner, J. P., Niskanen, A., and Järvisalo, M. (2017). Complexity Results and Algorithms for Extension Enforcement in Abstract Argumentation. *Journal of Artificial Intelligence Research (JAIR)*, 60:1–40.

[Winslett, 1990] Winslett, M. (1990). *Updating Logical Databases*. Cambridge University Press.

[Zheleznyakov et al., 2020] Zheleznyakov, D., Kharlamov, E., Nutt, W., and Calvanese, D. (2020). On Expansion and Contraction of DL-Lite Knowledge Bases. *CoRR*, abs/2001.09365.

[Zhuang and Pagnucco, 2014] Zhuang, Z. and Pagnucco, M. (2014). Entrenchment-Based Horn Contraction. *Journal of Artificial Intelligence Research (JAIR)*, 51:227–254.

[Zhuang et al., 2017] Zhuang, Z., Pagnucco, M., and Zhang, Y. (2017). Inter-Definability of Horn Contraction and Horn Revision. *Journal of Philosophical Logic*, 46(3):299–332.

[Zhuang et al., 2019] Zhuang, Z., Wang, Z., Wang, K., and Delgrande, J. P. (2019). A Generalisation of AGM Contraction and Revision to Fragments of First-Order Logic. *Journal of Artificial Intelligence Research (JAIR)*, 64:147–179.

[Zhuang et al., 2016] Zhuang, Z., Wang, Z., Wang, K., and Qi, G. (2016). DL-Lite Contraction and Revision. *Journal of Artificial Intelligence Research (JAIR)*, 56:329–378.

[Zhuang et al., 2013] Zhuang, Z. Q., Pagnucco, M., and Zhang, Y. (2013). Definability of Horn Revision from Horn Contraction. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, pages 1205–1212.

[Zwicker, 2016] Zwicker, W. S. (2016). Introduction to the Theory of Voting. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors, *Handbook of Computational Social Choice*, pages 23–56. Cambridge University Press.