# ABSTRACT

Title of Dissertation: DEFEASIBILITY IN EPISTEMOLOGY

Aleks Knoks
Doctor of Philosophy, 2020

Dissertation Directed by: Professor John Horty
Department of Philosophy &
Institute for Advanced Computer Studies

This dissertation explores some ways in which logics for defeasible reasoning can be applied to questions in epistemology. It's naturally thought of as developing four applications:

The first is concerned with simple epistemic rules, such as "If you perceives that $X$, then you ought to believe that $X$" and "If you have outstanding testimony that $X$, then you ought to believe that $X$." Anyone who thinks that such rules have a place in our accounts of epistemic normativity must explain what happens in cases where they come into conflict—such as one where you perceive a red object and are told that it is blue. The literature has gone in two directions: The first suggests that rules have built-in unless-clauses specifying the circumstances under which they fail to apply; the second that rules do not specify what attitudes you ought to have, but only what counts in favor or against having those attitudes. I express these two different ideas in a defeasible logic framework and demonstrate that there's a clear sense in which they are equivalent.

The second application uses a defeasible logic to solve an important puzzle about epistemic rationality, involving higher-order evidence, or, roughly, evidence about our capacities for evaluating evidence. My solution has some affinities with a certain popular view on epistemic dilemmas. The third application, then, is a characterization of this *conflicting-ideals view* in logical terms: I suggest that it should be thought of as an unconventional metaepistemological view, according to which epistemic requirements are not exceptionless, but defeasible and governed by a comparatively weak logic.

Finally, the fourth application is in the burgeoning debate about the epistemic significance of disagreement. The intuitive conciliatory views say, roughly, that you ought to become less confident in your take on some question $X$, if you learn that an epistemic equal disagrees with you about $X$. I propose to think of conciliationism as a defeasible reasoning policy, develop a mathematically precise model of it, and use it to solve one of the most pressing problems for conciliatory views: Given that there are disagreements about these views themselves, they can self-defeat and issue inconsistent recommendations.

# DEFEASIBILITY IN EPISTEMOLOGY

by

Aleks Knoks

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor John Horty, Chair/Advisor
Professor Fabrizio Cariani (Northwestern)
Professor Maria Lasonen-Aarnio (Helsinki)
Professor Eric Pacuit
Professor Paolo Santorio

# Acknowledgments

While working on the ideas laid out in this dissertation, I received a lot of support. It came in different forms and from different people, but it has been invariably important. First off, I have to thank the members of my advising committee Fabrizio Cariani, Maria Lasonen-Aarnio, Eric Pacuit, and Paolo Santorio. Their helpful advice and pointed questions have led to substantial improvements in the dissertation. Presenting my ideas to them in individual conversations has been memorable, fun, and, at times, scary. But the help that came from without the committee has also been very valuable. Here my thanks go to the participants of the conferences, workshops, and colloquiums in Amsterdam, Berlin, Cologne, College Park, Helsinki, Palo Alto, Providence, Rīga, Storrs, and Uppsala where my ideas where presented. Special thanks go to David Christensen, Benjamin Kiesewetter, Jim Pryor, and Thomas Schmidt for in-depth discussions, detailed written comments on parts of the dissertation, as well as encouragement that came at the right time.

Moving on to other forms of support, I would like to thank Aidan Lyon, Edgars Narkēvičs, and Frank Veltman who have left their marks on this dissertation by having had a lasting impact on my intellectual development. I am also extremely grateful to my best friends from graduate school Quinn Harr and Julius Schönherr. Their friendship and support was what kept me afloat during the most difficult moments. My family's support has also been tremendously important, and I am very grateful to Ainārs, Artūrs, Anna-Marija, my mom Ludmila, as well as Dace and my pride and joy Freya for being there for me. I am especially grateful to you,

Dace, for enduring your dissertating husband. I know how difficult it's been.

My biggest thanks goes to my advisor Jeff Horty who has shown me how to approach philosophical questions in a very special way, and who is the one person whose support has taken almost all of its possible forms.

While writing this dissertation, I benefited from a grant from the German Academic Exchange Service (DAAD), as well as the Ann G. Wylie Dissertation Fellowship. This support too is gratefully acknowledged.

Table of Contents

# List of Figures

# Introduction

This dissertation is an exercise in what one might call *nontraditional formal* epistemology. It is both common and natural to define formal epistemology in opposition to its more traditional cousin: Where traditional epistemology approaches (normative) questions related to belief, knowledge, and reasoning relying on the classical method of conceptual analysis, formal epistemology approaches these very same questions drawing on tools from mathematics and logic. In principle, epistemological questions could be tackled using various formal frameworks, but, as a matter of fact, they are usually tackled using one of the two popular ones. The first is the Bayesian framework, or, roughly, a combination of probability theory and inductive logic. And the second is normal modal logic, or modal logic that can be given a possible worlds semantics. These are the standard formal tools in epistemology, and it is actually not uncommon to think that formal epistemology just is the applications of one of them—especially, the Bayesian framework—to epistemological questions. Such applications, then, are what I think of as standard or *traditional formal* epistemology; and what's going on in this dissertation can be understood in opposition to it. For what I'll be concerned with throughout the text is tackling epistemological questions, drawing on an entirely different framework, namely, the logics for defeasible reasoning.

The first such logics were developed in the field of artificial intelligence in response to the challenge to represent the information that would let a machine exhibit

intelligent behavior. Efforts to meet this challenge quickly made it clear that ordinary logic is utterly inadequate for this task, as much of this information takes the form of defeasible generalizations. Think of statements like "Birds fly" and "Things that look red are red" which express sensible principles of reasoning—principles we'd seem to constantly rely on in our everyday life—even though they allow for exceptions. Defeasible logics, then, are logics of such defeasible generalizations, and their study has long since grown into a significant area of research, intersecting theoretical computer science and philosophical logic.[1] But, more importantly for our purposes, defeasible logics can also be applied to normative questions in philosophy. In fact, the overall goal of this dissertation is to make a case for the following thesis: Logics for defeasible reasoning can be of great help in answering questions about the structure of epistemic normativity and furthering important debates in epistemology.

My strategy for making this case is to develop three independent and equally important applications of defeasible logics to questions from epistemology. Accordingly, the dissertation is divided into three parts.

The first part aims at a better understanding of simple epistemic rules, such as the following two:

> (Perception) If a (rational) agent perceives that $X$, then she ought to believe that $X$; and
>
> (Testimony) If a (rational) agent has outstanding testimony that $X$, then she

[1] For a good introduction to the applications of (defeasible) logics to problems from artificial intelligence see (Thomason 2018).

ought to believe that $X$.

It is natural to think that rules like these have a role to play in our accounts of epistemic normativity. But anyone who thinks that Perception, Testimony, or other rules like them are genuine immediately faces the challenge of explaining what happens in cases of *epistemic conflict*, or situations where such rules support opposing conclusions. Just think of a case where an object in front of you looks red and an extremely reliable source tells you that this object is blue. Were we to apply the two rules in this case, we would seem to have to conclude that you ought to believe both that the object in front of you is red and that it is blue. There are two plausible strategies of response to this problem, and they both weaken the above statement of the rules. According to the first, simple epistemic rules have implicit hedges or unless-clauses that specify the circumstances under which the rule doesn't apply. According to the second strategy, the "ought" that occurs in the consequent of epistemic rules is to be substituted with "has reason".

These two views of simple rules are usually thought of and presented as being distinct. In the first part of this dissertation, I do two things. First, I express both ways of thinking about simple rules in a mathematically precise way, using a defeasible logic framework. Second, I show that these two seemingly different ways of thinking are, in fact, much closer than standardly thought. Indeed, there's a straightforward sense in which they are equivalent. And it's not only that the models of these views handle particular cases in the same ways. We can define certain notions familiar from traditional epistemology—including those of an epistemic reasons, defeat,

and defeater—in each of these models, and they come out corresponding one-to-one. The correspondence result isn't of theoretical significance only, but also has some far-reaching consequences for the two views on rules.

In the second part, the focus shifts from epistemic rules to epistemic requirements. It's naturally seen as doing two things. First off, I use defeasible logic to work out a solution to an important puzzle about epistemic rationality: In case one's (total) evidence can be misleading about what it itself supports—as some epistemologists have recently argued—then two intuitive and widely accepted epistemic requirements can come into conflict, suggesting that there are dilemmas of rationality. My defeasible logic-based solution has a number of attractive features when compared to the other solutions from the literature. However, it also comes with an unorthodox perspective on epistemic requirements, a perspective on which they are defeasible. I also show—and this is the second major idea of this part of the dissertation—that we can naturally make sense of defeasible epistemic requirements as (regulative) epistemic ideals, and that the defeasible logic used to solve the puzzle can be naturally seen as the formal backbone of the *conflicting-ideals view* that David Christensen (2007a, 2010a, 2013) has been advocating for in his recent work. In effect, I'm proposing to understand this view as a move away from the default metaepistemological position according to which epistemic requirements are strict and governed by a strong, but never explicitly stated logic, toward the more unconventional view, according to which requirements are defeasible and governed by a comparatively weak logic. When understood this way, the view is not committed to the existence of dilemmas.

In the third part, I apply the logics for defeasible reasoning in the context of the debate about the epistemic significance of disagreement. More specifically, I use them to get a better handle on conciliatory views on disagreement and the logical structure of conciliatory reasoning. On the view that emerges, conciliationism is a second-order defeasible reasoning policy, saying roughly the following: If your best (first-order) reasoning suggests that $X$ and it's rational for you to think that an epistemic peer disagrees with you regarding $X$, you should not conclude that $X$ under normal circumstances. Within the defeasible logics that I use to model conciliationism, the phrases "it's rational for you to think" and "under normal circumstances" have precise content. In the course of the three chapters that make up the third part, I do not only develop a precise model of conciliationism, but also use it to address a pressing—perhaps, the most pressing—challenge for conciliatory views: Given that there are disagreements about the epistemic significance of disagreement, conciliatory views can turn on themselves and—as Adam Elga (2010) has argued—thereby, also issue inconsistent directives.

The three parts are followed by an appendix that contains the proofs of all important observations.

# Part I DEFEASIBLE RULES

# Chapter 1: Two logics for defeasible rules

## 1.1 Epistemic rules and epistemic conflicts

Consider the following two rules:

> (Perception) If an agent's epistemic situation includes a perception that $X$, then the agent ought to believe that $X$.[1]
>
> (Testimony) If an agent's epistemic situation includes outstanding testimony that $X$, then the agent ought to believe that $X$.[2]

Both of these rules have received a fair amount of attention in the recent epistemology literature, and it is natural think that they—and, perhaps, also other *simple* rules like them—have a role to play in our accounts of epistemic normativity. But anyone who accepts that there are simple rules like these must also explain what happens in cases of *epistemic conflict*, or situations in which such rules either support conflicting conclusions, or get undermined. Suppose that an object in front of you looks red, but an extremely reliable source tells you that it is blue. Or suppose that someone you consider an authority in epistemology tells you that Testimony is false. What should you do in these cases? If we apply the rules in the first one,

[1] Cf. Boghossian (2017), Chisholm (1980), Huemer (2000), Pollock (1995); Pollock & Cruz (1999) Pryor (2000).

What's meant by an epistemic situation here? Well, it's natural to think that, whatever theory of epistemic normativity turns out to be correct, it is very likely to specify certain *descriptive* features of the agent's situation as relevant to determining which doxastic states the agent ought to have. These features may include the agent's evidence, facts about her condition, her past, or other kinds of facts. The totality of all of these normatively-relevant features is what I call the agent's *epistemic situation*—cf. (Titelbaum 2015, Sec. 2).

[2] Cf. Bradley (2019), Elga (2007, 2010), and Titelbaum (2015).

we are quickly lead to conclude that you ought to believe that the object is red and that it is blue! And if we apply Testimony in the second, we are lead to conclude that you ought to disbelieve it—which is at least somewhat odd, if we think that it is a genuine epistemic rule.[3]

Both of the two most plausible strategies of response to the problem weaken the above statement of rules, or suggest that such rules must be *defeasible*.[4] According to the first, Perception, Testimony, and other simple rules have implicit *hedges* or unless-clauses that state the circumstances under which the rule doesn't apply. One could then say that in the first problematic scenario the conflict between Perception and Testimony is only apparent, because, say, Testimony doesn't apply when you have perceptions to rely on. According to the second strategy, the "ought" that occurs in the consequents of simple rules is to be changed for "has a reason". The thought here is that simple rules (by themselves) do not specify what doxastic attitudes you are required to have, but only what counts *in favor* or against having them. Applying this *contributory-rules* strategy to the same scenario, we would say that there's indeed a conflict between Perception and Testimony, but that it is resolvable, because, say, your perception outweighs the testimony.[5]

[3]Two clarificatory notes here: First, notice that the example I'm using is a special case of a situation where a rule gets undermined, namely, a situation where the rule self-undermines—cf. (Titelbaum 2015, Sec. 4). Second, notice that the conclusion I draw depends on an inter-level coherence principle, saying that it is never epistemically permissible to believe that you ought to disbelieve $X$ and believe $X$ all the same. Given the goals of this part of the dissertation, nothing important hinges on us assuming that it is genuine. I'll discuss my take on this principle in Chapter 3.

[4]Here the term *defeasible* is used in a loose sense, getting at the intuitive idea that a rule may engender an ought in one situation and then fail to engender an ought in a situation that differs from the original one only by a margin.

[5]The authors who have pursued the first general strategy in epistemology include Bradley (2019), Elga (2010), and Titelbaum (2015), and those who have pursued it in the moral domain include Holton (2002) and Scanlon (2000). The list of authors who may have pursued something

We will state the two strategies precisely later. For now note two things about them. First, they are naturally thought of and typically presented as distinct.[6] And second, it's neither obvious, nor uncontroversial that either one of them succeeds in responding to the problem while retaining a conception of simple rules that's sufficiently close to the one we started with. To take one example, Darren Bradley (2019) has recently argued that the second strategy fails, and that the first one, once spelled out in full detail, reduces to an extreme version of particularism in epistemology.

This part of the dissertation has two goals: The first is to express both ways of thinking about rules in a mathematically precise way, using a simple formal framework motivated by the work from logics for defeasible reasoning. The second goal is to show that these two seemingly different ways of thinking about rules are much closer than is standardly thought. Indeed, there's a straightforward sense in which they are equivalent. As will become clear, this result is not of only theoretical significance, but also provides important important insights into the nature of both views on rules.

This chapter is structured as follows. Section 1.2 presents the basic formal concepts and formalizes the problem that epistemic conflicts give rise to. Sections

akin to the second strategy in epistemology includes Christensen (2007a, 2010a, 2013), Horty (2012), Pollock (1995); Pollock & Cruz (1999). The view is much more wide-spread in ethics, where it is associated with W. D. Ross (1930). (Such authors as Lance & Little (2007), McKeever & Ridge (2006), and Väyrenen (2009) appear to defend views that combine the two strategies.) The two strategies (and their combination) seem to me to exhaust the space of plausible responses to the problem that don't do away with simple rules as such. A third, much less plausible, alternative is to bite the bullet and accept ubiquitous existence of epistemic dilemmas, or situations in which the agent fails epistemically no matter what doxastic state she adopts. The author who is the most likely to sympathize with this response is Hughes (2017).

[6]See, for instance, (Bradley 2019) in epistemology and (Dancy 2004, Sec. 1.2) in ethics.

1.3 and 1.4 develop, respectively, a model of the hedged-rules view and a model of the contributory-rules view. The first section of the subsequent Chapter 2 will establish the central result—that the contributory-rules view is equivalent to a *restricted* version of the hedged-rules view—and make use of the result to show how the hedged-rules view can account for two phenomena it is standardly thought not to be able to handle. Throughout this discussion we will focus on cases where simple rules support conflicting conclusions, as opposed to the ones where rules get undermined. But the topic of undermining will take central stage in Section 2.2 where I will discuss the limits of the contributory-rules view and establish a further equivalence between the *full* hedged-rules view and a mixed view, according to which rules are both contributory and hedged.

## 1.2 The problem formalized

As our background, we assume the language of ordinary propositional logic with the standard connectives. The turnstile $\vdash$ will stand for classical logical consequence: Thus, where $X$ and $Y$ are propositional formulas, $X \vdash Y$ means that $Y$ is a classical consequence of $X$. For the sake of convenience and in order to avoid unnecessary clutter in our formalization of particular cases, we will assume that our background language allows for materially inconsistent atomic formulas, which will let us express statements, such as "The object in front of you is red" and "The object in front of you is blue," that can't jointly be true at the same time. Also, in order to have a more natural way of stating Perception and Testimony, we extend the language

with three designated predicates, $Perceive(\cdot)$, $Testimony(\cdot)$, and $Believe(\cdot)$. Not surprisingly, $Perceive(X)$ captures the idea that the agent has a perception that $X$; $Testimony(X)$ that the agent has outstanding testimony that $X$; and $Believe(X)$ that the agent believes that $X$. The reference to an agent is important. Even though we do not represent the agent in the language, all the formulas we'll encounter should be thought of as relativized to an agent situated in some epistemic situation. We will also make use of the customary deontic operator $\bigcirc(\cdot)$. A formula of the form $\bigcirc Believe(X)$, then, expresses the idea that the agent ought to, or is rationally required to, believe that $X$.[7] A few quick remarks on the sense of 'ought' that I have in mind here: It should be understood in the epistemic, rather than pragmatic or any other sense; it should be understood as relative to the agent's epistemic situation;[8] and it should be understood as all things considered, as opposed to *pro tanto*.

Now we turn to the question of how to express Perception, Testimony, and other simple rules in this language. One might be tempted to use material conditionals, formalizing them as $Perceive(X) \supset \bigcirc Believe(X)$ and $Testimony(X) \supset \bigcirc Believe(X)$, respectively. However, given our goals, it will be better to think of them by analogy with the inference rules of natural deduction systems—for, eventually, we will want to turn Perception and Testimony into defeasible rules. Just like the rule of conjunction elimination sanctions one to conclude $X$ (as well as $Y$) whenever one has been able to establish $X\&Y$, a simple epistemic rule can be thought

[7] Following the usual practice in epistemology, I use 'ought to' and 'rationally required to' as synonyms.

[8] Some epistemologists focus on the sense of the 'ought' in which you ought to believe $X$ only if it is true that $X$, but that's not the sense that I have in mind here. I'm interested in the sense of 'ought', according to which you ought to believe $X$ if it is justified in your epistemic situation—cf. Bradley (2019).

of as sanctioning drawing a certain type of conclusion whenever one's epistemic situation includes a certain type of feature. One standard way of stating conjunction elimination presents it in a tree-form, as follows:

$$\frac{X\&Y}{X} \qquad\qquad \frac{X\&Y}{Y}$$

And we can state Perception and Testimony in an analogous way:

$$\frac{Testimony(X)}{\bigcirc Believe(X)} \qquad\qquad \frac{Perceive(X)}{\bigcirc Believe(X)}$$

It's worth highlighting an important technical detail: Strictly speaking, what we have here—in the case of conjunction elimination, as well as Perception and Testimony—are not rules, but *rule schemas*. How are rules and schemas related? Well, on the one hand, rules result from instantiating rule schemas with particular propositional formulas. And on the other, rule schemas let one state the rules that share a common form in a concise way. It's fairly common to elliptically refer to rule schemas as *rules*—and I will occasionally do so too—but we shouldn't forget that rules and schemas are distinct notions that shouldn't get confounded.

In what follows, we will focus on rules more than on rule schemas, and we will use the letter $r$ (with subscripts) to denote rules. Also, for reasons of presentation, we will sometimes write rules not in their tree-form, but rather as pairs of formulas $\langle X, Y\rangle$ with the first element $X$ standing for the premise of the rule, and the second element $Y$ standing for its conclusion.[9] It will also be useful to introduce two functions, $Premise[\cdot]$ and $Conclusion[\cdot]$, for picking out the premise and

[9]Compare to the way rules are represented in input/output logic—see e.g., (Makinson & van der Torre 2000, 2001) and (Parent 2011).

conclusion of a given rule. Thus, if $r$ stands for the rule $\langle X, Y\rangle$, then $Premise[r]$ is the proposition $X$ and $Conclusion[r]$ is the proposition $Y$. The second function will also be lifted from individual rules to sets of rules: Where $\mathcal{R}$ is a set of rules, $Conclusion[\mathcal{R}]$ is the set containing the conclusions of all the rules in $\mathcal{R}$, or $Conclusion[\mathcal{R}] = \{Conclusion[r] : r \in \mathcal{R}\}$.

We'll represent epistemic situations as pairs of the form $\langle \mathcal{W}, \mathcal{R}\rangle$, referring to them as *epistemic contexts* and denoting them with the lowercase $c$ (again, with subscripts). Any epistemic context will contain two elements. The first element $\mathcal{W}$, called the *hard information*, is a set of ordinary propositional formulas, expressing the descriptive features of the situation. The second element $\mathcal{R}$ is a set of epistemic rules, or pairs of formulas of the form $\langle X, \bigcirc Believe(Y)\rangle$.

**Definition 1.1 (Epistemic contexts)** *An* epistemic context *$c$ is a structure of the form $\langle \mathcal{W}, \mathcal{R}\rangle$, where $\mathcal{W}$ is a set of propositional formulas and $\mathcal{R}$ is a set of epistemic rules.*

As our first illustration of this notion, consider the simple case where you're looking at an object in front of you and it looks red. Letting $R$ stands for the proposition that the object in front of you is red, we can encode this situation in the epistemic context $c_1 = \langle \mathcal{W}, \mathcal{R}\rangle$ where $\mathcal{W}$ contains the formula $Perceive(R)$, standing for the proposition that the object in front of you looks red, and $\mathcal{R}$ includes the rule $r_1 = \dfrac{Perceive(R)}{\bigcirc Believe(R)}$, saying that you ought to believe that the object is red if it looks red to you.

It's natural to think that one important task of epistemology consists in identifying the simple epistemic rules that would let us link the descriptive features of

any epistemic situation with the doxastic attitudes that the agent ought to have in that situation. With the formal notion of an epistemic context in hand, we can think of this task as follows: There is a(n infinite) set of contexts that capture various sorts of epistemic situations, and these contexts share a common set of rules $\mathcal{R}$. The epistemologist's task, then, is to figure out (i) what is the shape of the rules in $\mathcal{R}$ (or what is the form of the rule schemas) and (ii) what is the logic of interaction between these rules that we need, for it to be the case that the $\bigcirc$-formulas that follow from each context $c$ match what we intuitively think is rationally required from the agent in the epistemic situation that $c$ stands for.

One's first-pass hypothesis might be that the rules in $\mathcal{R}$ will be the simple rules of the sort we have discussed and that the logic of their interaction will be classical. The formal framework lets us state this hypothesis in a precise way. We start by introducing the notion of a *triggered rule*.

**Definition 1.2 (Triggered rules)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be an epistemic context. The rules from $\mathcal{R}$ that are* triggered *in $c$ are those that belong to the set $Triggered(c) = \{r \in \mathcal{R} : \mathcal{W} \vdash Premise(r)\}$.*

So the rules that are triggered in a context are all and only those rules whose premises can be derived from $\mathcal{W}$ by classical logic. It's easy to see that $r_1$ is triggered in the context $c_1$. Since $Perceive(R)$ is in $\mathcal{W}$ and $Premise[r_1] = Perceive(R)$, we have $\mathcal{W} \vdash Perceive(R)$.

But we still need to specify how to get to the $\bigcirc$-formulas. There are a few different ways to do this, but, given our goals, we can simply adopt the most straight-

forward one, taking the conclusions of all the rules that are triggered in the context:[10]

**Definition 1.3 (Consequence, first pass)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be an epistemic context. Then the ought statement*

$$\bigcirc X \text{ follows from } c \text{ if and only if } \bigcirc X \in Conclusion[Triggered(c)].$$

Applying this definition to our example, it's very easy to see that the statement $\bigcirc Believe(R)$ follows from $c_1$. Since $r_1$ is the only rule triggered in $c_1$, the set $Conclusion[Triggered(c_1)]$ equals $\{\bigcirc Believe(R)\}$. And it's only one step from here to see that $\bigcirc Believe(R)$ follows from $c_1$, as desired.

At this point already we could state the problem that epistemic conflicts give rise to. But before we do that, it'll be useful to introduce another formal notion, namely, that of a *contrary rule*.

**Definition 1.4 (Contrary rules)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be an epistemic context and $r = \dfrac{X}{\bigcirc Believe(Y)}$ and $r' = \dfrac{Z}{\bigcirc Believe(W)}$ two rules from $\mathcal{R}$. Then the rules $r$ and $r'$ are* contrary in the context *$c$, written as $contrary_c(r, r')$, if and only if $Y$ and $W$ are inconsistent.*

Now recall the troubling epistemic situation we started with where an object in front you looks red and an extremely reliable source tells you that it is blue. Let $R$ be as before and let $B$ stand for the proposition that the object in front of you is blue. We can, then, encode this situation in the context $c_2 = \langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W}$

[10] I've chosen this definition to keep the formalism as simple as possible. Were the setting more formally-oriented, we could've, for instance, used the following definition: $\bigcirc X$ follows from $\langle \mathcal{W}, \mathcal{R} \rangle$ if and only if $\bigcirc X$ follows from $Conclusion[Triggered(c)]$, given *standard deontic logic*. For a nice presentation of this logic see (McNamara 2019, Sec. 2).

containing $Perceive(R)$ and $Testimony(B)$, and $\mathcal{R}$ containing $r_1 = \dfrac{Perceive(R)}{\bigcirc Believe(R)}$ and $r_2 = \dfrac{Testimony(B)}{\bigcirc Believe(B)}$. Note that the latter rule says that you ought to believe that the object is blue if you have outstanding testimony that it is blue. It's easy to see that $r_1$ and $r_2$ are both triggered in $c_2$, and that both $\bigcirc Believe(R)$ and $\bigcirc Believe(B)$ follow from $c_2$, suggesting that you ought to believe that the object is red and that it is blue. This, of course, is preposterous. Notice too that $r_1$ and $r_2$ qualify as contrary rules—in the end, $R$ and $B$ are materially inconsistent. This latter fact, in turn, point in the direction of a more general statement of the problem: In any context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ where two or more contrary rules from $\mathcal{R}$ get triggered, the agent will be required to have inconsistent beliefs.

In order to respond to this problem, we need to change either the way we think about simple rules, or the logic that lets lets us derive the $\bigcirc$-formulas, or both.

## 1.3 Hedged-rules strategy

### 1.3.1 What type of hedges?

The first general strategy of response suggests that simple rules have built-in hedges or unless-clauses specifying the circumstances under which the rules don't apply. The literature has taken this strategy in two different directions: One idea is that the content of rule hedges is *normative*, the other that it is *descriptive*.

The first idea is exemplified by Elga (2010) and Mike Titelbaum (2015), both of whom developed their views responding to cases where Testimony gets (self-)undermined, as opposed to cases where rules support contrary conclusions. On

Titelbaum's view, in particular, the hedge of a genuine epistemic rule must refer to "truth about what rationality requires", and the rule Testimony must be replaced by Testimony*:

> (Testimony*) If an agent's situation includes outstanding testimony that $X$, then the agent ought to believe that $X$—unless $X$ contradicts an *a priori* truth about what rationality requires, or, simply, unless $\bigcirc\neg Believe(X)$.[11]

Notice how the unless-clause helps with the problematic cases of self-undermining:[12] Supposing that Testimony* is a genuine epistemic rule, and that rationality requires that you believe it, Testimony* will simply *not* apply to any (misleading) testimony against it. Notice too that this wouldn't be particularly useful if the agent had no way of telling whether or not rationality does indeed require that she believes Testimony*. But on Titelbaum's view, this is something the agent can establish *a priori*:

> [E]very agent possesses *a priori*, propositional justification for true beliefs about the requirements of rationality in her current situation. An agent can reflect on her situation and come to recognize facts about what that situation rationally requires. Not only does this reflection provide her with justification to believe those facts; that justification is ultimately empirically indefeasible (Titelbaum 2015, p. 276).

Titelbaum's view is certainly interesting, but many epistemologists have pushed back against it, mostly on the basis of its extremely counterintuitive consequences.

[11] Compare to the statement of "Properly Restricted Testimony" on p. 274 of (Titelbaum 2015).
[12] In fact, the clause does much more. In effect, it makes the rule impossible to undermine.

For instance, the view would have it that a self-conscious fallible agent is to retain full confidence in Testimony* under all circumstances, no matter how much empirical evidence against it she may have.[13] What's more, independently of the view's success as a response to cases of undermining, it is of no help as a response to the problem of conflicts between rules. So this first way of thinking about rule hedges doesn't seem to hold much promise, and we will not discuss it any further.

On the second way of thinking, the hedges of simple rules refer to the descriptive features of the agent's situation. It appears that Bradley (2019) is the only one to systematically develop this idea in the context of epistemology, but it has been developed earlier in the ethics literature.[14] To see how this idea could help, we turn to a simple example. Consider an epistemic situation where a reliable source tells you that some object is red and another, equally reliable, source tells you that the same object is blue. Applying Testimony in this situation quickly leads to the conclusion that you ought to believe that the object is red and that is blue. But suppose that we supplement Testimony with a hedge, as follows:

> (Hedged Testimony) If an agent's epistemic situation includes outstanding testimony that $X$, then the agent ought to believe that $X$—unless the situation also includes testimony that is contrary to $X$ and at least as good.

This rule fails to apply in your situation. Why? Well, you do have an outstanding testimony that the object is red, but the conditions that are specified in the hedge are satisfied too: Your situation *also* includes an equally good testimony to the

[13] See e.g., (Christensen 2013, pp. 88–9) and (Bradley 2019, pp. 4–7).

[14] See footnote 5 for references. We will touch on hedged rules in ethics in Sections 2.1.2–2.1.3.

contrary. Consequently, it's not the case that you ought to believe that the object is red. And parallel reasoning applies to the testimony that the object is blue.

Now let's try capture this idea in the formal framework.

### 1.3.2 Strategy formalized

The first step is to change the notion of a rule. We have been thinking of simple epistemic rules as *ordered pairs* of formulas of the form $\langle X, \bigcirc Believe(Y)\rangle$. And it's fairly natural to think of hedged epistemic rules as *ordered triples* of formulas of the form $\langle X, \bigcirc Believe(Y), \mathcal{Z}\rangle$, where the first two element, $X$ and $\bigcirc Believe(Y)$, are still, respectively, the rule's premise and conclusion, and the third element $\mathcal{Z}$ is the rule's hedge. There's more than one way of specifying the shape of the hedge, but we will require that it is a set of negated formulas, or that $\mathcal{Z}$ is of the form $\{\neg Z_1, \ldots, \neg Z_n\}$. Hedged rules too can be represented in a tree form, as follows:

$$\frac{X : \neg Z_1, \ldots, \neg Z_n}{Y} .$$

A hedged rule of this form should be red as, "If $X$ obtains, then conclude $Y$, unless either $Z_1$, or ..., or $Z_n$ obtain". Or, alternatively, it can be read as, "If $X$ obtains, and it can be assumed that not-$Z_1$ and ..., and not-$Z_n$, then conclude $Y$."[15]

We will retain the functions for selecting rule premises and conclusions: Where $r$ stands for the hedged rule $\langle X, Y, \mathcal{Z}\rangle$, $Premise[r]$ is the proposition $X$ and $Conclusion[r]$ is the proposition $Y$. In addition, we will introduce another function to have access to rule hedges. Since sometimes this function will get applied to hedgeless rules, its

[15]Compare with Reiter's (1980) default rules.

definition is bipartite:

$$Hedge[r] \;=\; \begin{cases} \mathcal{Z} & \text{if } r \text{ is a hedged rule of the form } \langle X, Y, \mathcal{Z}\rangle \\ \varnothing & \text{otherwise, that is, if } r \text{ is of the form } \langle X, Y\rangle \end{cases}$$

So, upon being given a rule, the function outputs its hedge, if the rule has one, and otherwise it outputs the empty set. Above, we used the (formal) notion of a context to capture epistemic situations. Now we will capture such situations using *hedged contexts*, or contexts whose sets of rules can, and normally will, contain hedged rules.

**Definition 1.5 (Hedged epistemic contexts)** *A* hedged epistemic context *is a structure of the form* $\langle \mathcal{W}, \mathcal{R}\rangle$*, where* $\mathcal{W}$ *is a set of propositional formulas and* $\mathcal{R}$ *is a set of epistemic rules, possibly hedged.*[16]

Now consider a scenario where one reliable source tells you that the object in front of you is red and another reliable source tells you that that very object is blue. We can encode this scenario in a hedged context $c_3 = \langle \mathcal{W}, \mathcal{R}\rangle$ where $\mathcal{W}$ consists of $Testimony(R)$ and $Testimony(B)$, while $\mathcal{R}$ includes the hedged rules

$$r_3 = \frac{Testimony(R) \,:\, \neg Testimony(B)}{\bigcirc Believe(R)} \quad \text{and}$$

$$r_4 = \frac{Testimony(B) \,:\, \neg Testimony(R)}{\bigcirc Believe(B)}\,.$$

[16] A technical note: It's in general possible for a context to contain two rules whose premises and conclusions are the same, but hedges different. Intuitively, however, any pair of such rules is deviant—instead of two rules with different hedges, there should be only one rule specifying all the conditions under which the rule fails to apply. I will assume that hedged contexts never contain such deviant pairs of rules, or that the definition of a hedged context is subject to the following constraint: For any $r, r' \in \mathcal{R}$, in case $Premise[r] = Premise[r']$ and $Conclusion[r] = Conclusion[r']$, then $r = r'$.

Of course, we still need to specify how to derive $\bigcirc$-formulas from hedged contexts. When working with hedgless rules, we simply collected the conclusions of all the rules triggered by the context's hard information. Now we add an extra condition: The hard information must not only trigger the rule, but it must also *not* exemplify any of the features listed in the rule's hedge. The notion of an *admissible rule* captures both conditions.

**Definition 1.6 (Admissible rules)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a hedged context. The rules from $\mathcal{R}$ that are* admissible *in $c$ are those that belong to the set*

$$Admissible(c) = \{r \in \mathcal{R} : r \in Triggered(c) \text{ and, for no } \neg Z \in Hedge[r],\ \mathcal{W} \vdash Z\}.$$

Now let's apply the notion to the context $c_3$. It's not hard to see that neither the rule $r_3$, nor $r_4$ come out admissible. Why? Well, take $r_3$: Its premise $Testimony(R)$ does follow from $\mathcal{W}$, and so it qualifies as triggered. However, its hedge contains the formula $\neg Testimony(B)$ and $Testimony(B)$ follows from $\mathcal{W}$. Analogous reasoning applies to $r_4$.

With the notion of an admissible rule in hand, all we need to do to specify how to get the $\bigcirc$-formulas from a context is substitute $Admissible(c)$ for $Triggered(c)$ in Definition 1.3 from the previous section.

**Definition 1.7 (Consequence, hedged)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a hedged epistemic context. Then the ought statement*

$$\bigcirc X \text{ follows from } c \text{ just in case } \bigcirc X \in Conclusion[Admissible(c)].$$

It's straightforward to see that neither $\bigcirc Believe(R)$, nor $\bigcirc Believe(B)$ follow from $c_3$, as desired. So far, so good. However, part of the problem with epistemic conflicts

is that we always seem to be able to come up with further cases of conflicts between rules. Thus, suppose that, right after you receive the two conflicting testimonies, you take a look at the object yourself, and find out that it looks red to you. Intuitively, under these circumstances, you ought to believe that the object is red. The new situation can be captured using the context $c_4 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ contains the propositions $Testimony(R)$, $Testimony(B)$, and $Perceive(R)$, while $\mathcal{R}$ includes two rules we have encountered before, $r_1$ and $r_3$, as well as an updated version of $r_4$, which we refer to as $r_4'$:

$$r_1 = \frac{Perceive(R)}{\bigcirc Believe(R)},$$

$$r_3 = \frac{Testimony(R) \,:\, \neg Testimony(B)}{\bigcirc Believe(R)}, \text{ and}$$

$$r_4' = \frac{Testimony(B) \,:\, \neg Testimony(R), \neg Perceive(R)}{\bigcirc Believe(B)}.$$

A routine check will convince you that the only rule that's admissible in this context is $r_1$. So, $\bigcirc Believe(R)$ does, while $\bigcirc Believe(B)$ does not follow from $c_4$. There's a natural question about the relation between $r_4$ and $r_4'$ that you might have at this point: Should we say that the hedge of the simpler rule changes once you look at the object, or should we, rather, say that $r_4$ was the rule $r_4'$ all along? Let's bracket this question for a second and think about the way this epistemic vignette could develop further.

Suppose that, after you have taken a look at the object, the Epistemology Oracle approaches you and tells you that the object is in fact blue.[17] Surely, now the right thing for you to do is to believe what the oracle says! We will capture this

[17] I learned about the existence of epistemology oracles from White (2005).

latest (and final) development of the situation using the context $c_5 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ is as it was in $c_4$, except for now it also includes the formula $Testimony_O(B)$, standing for the Oracle's testimony, and the set of $\mathcal{R}$ includes the following rules:

$$r_1' = \frac{Perceive(R) : \neg Testimony_O(B)}{\bigcirc Believe(R)} ,$$

$$r_3' = \frac{Testimony(R) : \neg Testimony(B), \neg Testimony_O(B)}{\bigcirc Believe(R)} ,$$

$$r_4' = \frac{Testimony(B) : \neg Testimony(R), \neg Perceive(R)}{\bigcirc Believe(B)} , \text{ and}$$

$$r_5 = \frac{Testimony_O(B)}{\bigcirc Believe(B)} .$$

Notice the changes in the above rules $r_1$ and $r_3$. First, the instance of Perception $r_1'$ now too has a hedge that makes it inapplicable in any situation where $Testimony_O(B)$ obtains. Second, the hedge of the rule $r_3'$ now also includes the formula $\neg Testimony_O(B)$. This should make good sense. The Epistemology Oracle's testimony that the object is blue provides for a type of circumstance in which the original instance of Testimony shouldn't apply, independently of whether or not the original contrary testimony $Testimony(B)$ is present. Let's quickly convince ourselves that the above definition of consequence gives us the intuitively correct result: There's only one rule that's admissible in the context $c_5$, namely, $r_5$. The other three rules are all triggered, but they all are precluded from being added to the set of admissible rules because of their hedges. As a result, only $\bigcirc Believe(B)$ follows from the context, as desired.

The sequence of cases we have discussed motivates the following thought. The hedge of any given rule from $\mathcal{R}$ is actually not the simple set that I have presented

it to be, but, rather, a much more extensive one, a set that specifies all the possible circumstances under which the rule would fail to apply, most of which we haven't even conceived of yet, let alone discussed. I think this idea is correct, and that it is a natural consequence of the strategy we're pursuing. The one author who develops the hedged-rules view in epistemology, Bradley (2019), doesn't only think so too, but also takes this to mean that the view reduces to a position according to which there is but one unwieldy, incredibly complex, and not finitely expressible "Über-rule" (or, to put it in different terms, that the view reduces to a version of particularism in epistemology). What this Über-rule does, by stipulation, is specify the appropriate doxastic responses for all the epistemic situations that one might find oneself in.[18] The main reason for Bradley's suggestion that the hedged-rules view may reduce to an Über-rule view appears to be the prospect of rule hedges being "open-ended", as opposed to containing a finite number of exceptions. Thus, he writes:

> Starting with simple rules, can the exceptions be finitely stated? Ideally, we would like to have finite exceptions, as this would allow a manageable set of rules that could be used to guide our deliberation. [..] I don't know if this is possible, so I will concede the point, and defend the possibility that the exceptions are open-ended (Bradley 2019, p. 12).

How shall we make sense of this open-endedness? Well, here I think we have to think back to the distinction between rules and rule schemas, and to make sense of the question that Bradley is posing as a question about schemas, not rules. Take

[18]For discussion of the Über-rule see (Bradley 2019, Sec. 8), (Christensen 2010a), (Christensen 2013), and (Lasonen-Aarnio 2014).

some simple rule schema, such as Testimony. Now refocus on its instances which, let's suppose, you take to be hedged rules. Then there are two options. In case it turns out that there's a finite number of *types* of features that occur in the hedges of these rules, we will be able to write down an informative and, in the best case scenario, also usable schema covering all of them. For instance, all of the instances of Testimony we have encountered so far can be captured by the following schema:

$$\frac{Testimony(X) : \neg BetterTestimony(Y), \neg Perception(Y)}{\bigcirc Believe(X)},$$

with the background assumptions that it would be irrational for the agent to believe both $X$ and $Y$, and that $BetterTestimony(Y)$ expresses the intuitive thought that the agent's testimony for $Y$ is at least as good as her testimony for $X$. (Let me emphasize that I'm not suggesting that this is the true Hedged Testimony schema, but only that it captures all the cases we have discussed. The true schema would be more complex, but still have the same general shape.)

If, on the other hand, it turns out that there are infinitely many *types* of features that occur in the hedges of instances of Testimony, if it turns out that the circumstances under which these rules fail to apply are utterly different, then we will not be able to write down a rule schema that would cover all of them and be as informative as the one above. For were we to opt for informativeness, taking into account all of the rules, we'd seem to have no other option but listing these rules one by one—rules of which there's infinitely many. And were we to opt for a schema that can actually be written down, we would end up with something along the lines of, "If an agent's epistemic situations includes an outstanding testimony that $X$,

then she ought to believe that $X$—unless something comes in the way".

While there's something about the first schema that makes it seem intuitively preferable to the open-ended one, I am not quite sure what exactly is wrong with the latter. I am also not sure why the view on which hedged rule schemas are open-ended reduces to the Über-rule view, as Bradley suggests.[19] Nevertheless, I am going to concede the point and suppose that it does. In the end, this question—as well as the question of whether or not a view on which hedges are open-ended is a successful response to the problem posed by epistemic conflicts—is orthogonal to the goals I'm pursuing in this chapter. We will formulate a version of the hedged-rules view on which hedges contain a finite number of exceptions in Section 2.1 and revisit the question again in Section 2.2.2.

### 1.3.3 Reasons, defeat, and defeaters

Having developed a feel for the way hedged rules and context function, we turn to the question of how to think about some familiar epistemological concepts in the formal model, most notably, those of *epistemic reason*, *defeat*, and *defeater*. This is important because it will help us highlight the close connections between the hedged-rules view and the contributory-rules view—which we will turn to in Section

[19] As the above quote illustrates, Bradley concedes the possibility of open-ended hedges. And since he goes on to counter arguments against the Über-rule view, it's clear that he takes open-endedness to be tantamount to postulating such a rule. But why this is the case is never really made explicit. The closest we get to an explanation is the suggestion that, "Th[e Über-rule] position can be generated by conjoining an infinite number of simple rules or positing a finite number of simple rules with at least one infinitely long hedge" (Bradley 2019, p. 13)—cf. Holton (2002). I suspect that Bradley is motivated by the lack of a firm grasp on the shape that the hedged-rules view takes, once we allow that rule hedges are open-ended: Let them be open-ended, and it might just as well be that you end up with some rules whose hedges are as complex as the Über-rule is.

1.4.[20]

It should be clear that, in the model, the hedged rules connecting the descriptive and the normative are to be taken as primitive, and that other notions are to be defined in their terms. Let's start with reasons. Suppose we have an agent who finds herself in a nontrivial epistemic situation that's captured using some hedged context $c = \langle \mathcal{W}, \mathcal{R} \rangle$. What are the epistemic reasons that the agent has in this situation? Now, my proposal is that we identify them with the premises of those hedged rules that are triggered in this context. More specifically, the proposition $X$ is to be defined as a reason for $Y$, in the context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ if an only if there is a rule $r = \dfrac{X : \neg Z_1, \cdots, \neg Z_n}{\bigcirc Y}$ in $\mathcal{R}$ such that $r \in \mathit{Triggered}(c)$. In this case, we can say that $r$ *provides* a reason for $Y$, that $X$'s being a reason for $Y$ *depends* on $r$, and that $X$ and $Y$ stand in the *reason relation*. To see how this proposal works in a concrete situation, we'll look at the hedged context $c_5 = \langle \mathcal{W}, \mathcal{R} \rangle$ which we used to capture the final development of our three-part epistemic vignette. All of the four rules from $\mathcal{R}$ we discussed, namely, $r_1$, $r_3'$, $r_4'$, and $r_5$, came out triggered in the context of $c_5$. So, on the proposal, all of their premises are to be qualified as epistemic reasons that the agent has. And more specifically, the propositions that you have an outstanding testimony that the object is red, $\mathit{Testimony}(R)$, and that you perceive the object to be red, $\mathit{Perceive}(R)$, are your reasons for believing that it is red, and the propositions that you have an outstanding testimony that the object is blue, $\mathit{Testimony}(B)$, and that you have the Oracle's testimony that the object is blue,

[20] Although we will highlight these connections using the concept of a reason, we could do the same focusing on other concepts, such as *epistemic justification*.

$Testimony_O(B)$, are your reasons to believe that the object is blue.

Next question: How are we to make sense of defeat and defeaters? It is both natural and standard to have them related to the notions we have just defined, namely, reasons and the reason relation.[21] However, given that in the present framework rules are basic and reasons are made sense in their terms, we will be thinking of defeat as a notions that, in the first instance, too applies to rules. As before, let us fix a context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ and zoom in on some rule $r$ from $\mathcal{R}$. We will say that $r$ is *defeated* when it is among the rules that are triggered but not admissible in $c$. In other words, $r$ is defeated in $c$ if and only $r$ is in $Triggered(c)$ and there is a $\neg Z$ in $Hedge[r]$ such that $\mathcal{W} \vdash Z$. Not surprisingly, we will say that the formula $Z$ is a *defeater* of $r$. How does this play out at the level of reasons? Well, suppose that some proposition $X$ is a reason for $Y$, and that $X$'s being a reason for $Y$ depends on the hedged rule $r$. We would say that $X$ is defeated as a reason for $Y$ by a consideration $Z$ when $r$ is defeated by $Z$.

To have a concrete illustration, let's take a look at the context $c_5$ again. We have noted already that the rule

$$r'_4 = \frac{Testimony(B) \,:\, \neg Testimony(R),\, \neg Perceive(R)}{\bigcirc Believe(B)}$$

is triggered in $c_5$ and that $Testimony(B)$ being a reason for believing that $B$ depends on it. Since $r'_4$ is not among the rules that are admissible in $c_5$, it is defeated in $c_5$.

[21] For instance, Pollock (1974) in his classic treatment works with his technical notion of a prima facie reason for belief: To say that $X$ is a prima facie reason for some agent to believe that $Y$ is to say that in the absence of any other information the agent would be justified in believing that it wouldn't be true that $X$ unless it were true that $Y$. A defeater, then, is a prima facie reason for thinking that this is not the case, or that it would not be true that $X$ unless it were also true that $Y$—see (Pollock 1974, pp. 41–2).

But which consideration is it defeated by? Well, it actually happens to have two defeaters! Both $Testimony(R)$ and $Perceive(R)$ are in the hard information $\mathcal{W}$ of the context, and both of them are listed in the rule's hedge. So $r_4'$ is defeated by $Testimony(R)$, as well as by $Perceive(R)$.

So our model of the hedged-rules view is fairly simple, but it is expressive enough to give a precise characterization of some central epistemological concepts. We could use it to capture further concepts too, and we will use it when thinking about the well-established distinction between *rebutting* and *undermining* defeat in Section 2.2. But for now, we have defined all of the concepts we need, and we can turn to the competitor of the hedged-rules view. Before we do that, however, it will be worthwhile to mention an alternative way of thinking about hedges to forestall a potential objection.

### 1.3.4 Digression: An alternative way to think about hedges

It's possible that at this point you think that our model of the hedged-rules view is unnecessarily complex, and that we could have, in fact, developed a simpler model, a model that doesn't retreat from the classical logic to a defeasible one, had we only expressed hedged rules in a different way. More specifically, you may think that I should have formulated such rules not as triples of the form $\langle X, Y, \mathcal{Z}\rangle$, but, rather, as pairs of formulas with the first element specifying both the premise of the rule and its hedge, and the second element specifying its conclusion. Let's consider an example to make this line of thought more concrete. One of the rules that we

needed to account for the three-episode vignette was:

$$r_3' = \frac{Testimony(R) : \neg Testimony(B), \neg Testimony_O(B)}{\bigcirc Believe(R)}.$$

The idea, then, is that, instead of $r_3'$, we should have gone for:

$$r_6 = \frac{Testimony(R) \& \neg Testimony(B) \& \neg Testimony_O(B)}{\bigcirc Believe(R)}.$$

A clear benefit of using $r_6$ over $r_3'$, so the thought goes, is that we don't even need a fancy logic that can handle hedges, and that classical logic alone suffices for generating intuitively appropriate $\bigcirc$-formulas from contexts.

Unfortunately, this line of reasoning is wrong. Admittedly, there's something that $r_6$ with classical logic in the background get right, namely, that $\bigcirc Believe(R)$ will not follow in the contexts where it shouldn't, namely, those that include the formulas $Testimony(B)$ and $Testimony_O(B)$. But the problem is that $\bigcirc Believe(R)$ will *not* follow in those contexts where it should follow either. To see this consider the epistemic situation where you're given only one piece of information, namely, that the object in front of you looks red. We can encode it in the (unhedged) context $c_6 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ is comprised of $Testimony(R)$ and $\mathcal{R}$ includes the rule $r_6$. Now note that the rule $r_6$ alone doesn't let us get the formula $\bigcirc Believe(R)$ from $Testimony(R)$. Why? Well, for the rule to get triggered, that is, for it to be the case that $\mathcal{W} \vdash Premise[r_6]$, the context's hard information $\mathcal{W}$ would also need to include $\neg Testimony(B)$ and $\neg Testimony_O(B)$. But it doesn't.

There is a way, familiar from the defeasible logics literature, to force the rule $r_6$ to get triggered in the context $c_6$, as well as all the other contexts that contain

$Testimony(R)$, but neither $Testimony(B)$, nor $Testimony_O(B)$. Described at a high level, the trick is to treat all the formulas as false, unless specified otherwise. Were we to make use of this trick, upon receiving $c_6$, the formalism would treat $Testimony(B)$ and $Testimony_O(B)$ as false, or, as it were, add $\neg Testimony(B)$ and $\neg Testimony_O(B)$ to $\mathcal{W}$; and this would allow the rule $r_6$ to get triggered. In the literature, this is know as the *closed-world assumption* or *closed-world reasoning*.[22] However, making use of this trick also means departing from classical logic and embracing a defeasible consequence relations. So defeasible logics appears to be inseparable from the hedged-rules view.

One last question: Why opt for a view on rules where hedges and premises are kept distinct, as opposed to one where they aren't and supplementing it with the closed-world assumption? The answer is simple: The former makes for a more nuanced and more expressive view. (This will become clear in Section 2.1)

## 1.4 Contributory-rules strategy

### 1.4.1 General strategy and the model

The second general strategy for responding to the problem that epistemic conflicts give rise suggests that simple epistemic rules are, in fact, contributory. On this proposal, Testimony doesn't say that you ought to believe that $X$ in any epistemic situation that includes outstanding testimony that $X$, but only that you *have a reason* to believe that $X$ in any such situation. How does this help with the

[22] See (Reiter 1978).

problem? Well, suppose that you're in a situation where the object looks red to you, but a reliable source tells you that it is blue. If Perception and Testimony are contributory, we only get the conclusions that you have a reason to believe that the object is red and that you have a reason to believe that the object is blue. Reasons, of course, are *pro tanto*, and so a dilemmic conflict gets reduced to a conflict that we can live with.

But how are we to capture contributory rules, and the corresponding view, in the formalism? Let's start with Contributory Perception. One way of expressing it is:

$$\frac{Perceive(X)}{\text{There's a reason to } Believe(X)}\ .$$

Now, if all one cares about is avoiding the problem, then this schema certainly does the job, as "There's a reason to $Believe(X)$" is never going to be in contradiction with any statement of the form "There's a reason to $Believe(Y)$," even if $X$ and $Y$ are inconsistent. But there's also a good reason to be dissatisfied with this schema. Recall that simple rules were supposed to get us from a description of any particular epistemic situation to the doxastic attitudes that the agent ought to have in that situation. In our formal setting, this is the question of which $\bigcirc$-formulas follow from a given epistemic context. The problem is that it isn't clear how we are to get to the statement $\bigcirc Believe(X)$, or, perhaps, $\bigcirc Believe(Y)$, from "There's a reason to $Believe(X)$" and "There's a reason to $Believe(Y)$." Here one might, of course, say that this will depend on the relative weights of the relevant reasons, and that's exactly right. However, the suggestion doesn't actually bring us much to deriving

◯-formulas from contexts. (Note too that a statement of the form "There's a reason to $Believe(X)$" doesn't even tell us what the reason for believing $X$ is.)

Luckily, there's a better way to capture contributory rules. Taking a cue from John Horty (2012) and John Pollock (1995, 1999), we can think of Contributory Testimony and Contributory Perception as the following two *defeasible* rule schemas:

$$\frac{Testimony(X)}{Believe(X)} \qquad\qquad \frac{Perceive(X)}{Believe(X)} \ .$$

How are we to understand instances of these schemas? Well, let's zoom in on one such, the rule $r_7 = \dfrac{Testimony(R)}{Believe(R)}$. Intuitively, $r_7$ should be understood as saying that $Testimony(R)$ *favors* or *counts in favor of* believing $R$, or that the testimony that the object in front of you is red counts in favor of believing that it is red. Note that there's no presumption that the epistemic situation, against the background of which we are thinking about $r_7$, must include $Testimony(R)$. Rather, the favoring relation should be thought of as hypothetical: If $Testimony(R)$ obtained in the situation, then it would favor believing $R$. Functionally, what $r_7$ does for us in the model is let us infer $Believe(R)$ once $Testimony(R)$ has been established *by default*. The qualification "by default" is very important, and it is added because the presence of $Testimony(R)$ in an epistemic situation does not yet guarantee that it will be possible to infer $Believe(R)$, as other rules might come in the way. (We will see how this works in detail in just a little bit.) A huge advantage of thinking of contributory rules in this way is that there's a method for deriving ◯-formulas from such rules and their interaction within a hand's reach.

It's both standard and natural to associate contributory rules with relative

weights—in the end, they express the favoring relations, and everyone agrees that those come with relative weights.

To capture these weights in the formalism, we will introduce a new device: a *priority relation* over rules.[23] Where $r$ and $r'$ are simple (unhedged) epistemic rules, a statement of the form $r \leq r'$ will mean that $r'$ is at least as strong as $r$, or that $r'$ has at least as much weight as $r$. We will require that the relation $\leq$ satisfies some natural properties. First, it must satisfy the *reflexivity* property,

$$r \leq r',$$

according to which each rule is at least as strong as itself. Second, the relation $\leq$ must also satisfy the *transitivity* property,

$$r \leq r' \text{ and } r' \leq r'' \text{ entail } r \leq r'',$$

according to which whenever $r''$ is at least as strong as $r'$ and $r'$ is at least as strong as $r$, $r''$ must be at least as strong as $r$. Any relation satisfying reflexivity and transitivity is called a *preorder*, and so I will sometimes refer to $\leq$ as a preorder. It will also be useful to introduce some shorthand: When we have $r \leq r'$ without $r' \leq r$, we will write $r < r'$. And when we have both $r \leq r'$ and $r' \leq r$, we will write $r \sim r'$.

In Sections 1.2 and 1.3, we captured epistemic situations using the notion of a context, which was always a pair of the form $\langle \mathcal{W}, \mathcal{R} \rangle$. Here we will capture epistemic situations using the notion of a *weighted context* that adds a third element to the pair, namely, a preorder $\leq$ on the rules in $\mathcal{R}$.

[23] This is a standard move—see e.g., (Pollock 1995) and (Horty 2012).

**Definition 1.8 (Weighted epistemic contexts)** *A weighted epistemic context $c$ is a structure of the form $\langle \mathcal{W}, \mathcal{R}, \leq \rangle$, where $\mathcal{W}$ is a set of ordinary propositional formulas, $\mathcal{R}$ is a set of contributory rules, and $\leq$ is a preorder on $\mathcal{R}$.*[24]

Our first weighted context $c_7 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ will capture the familiar epistemic situation where one reliable source tells you that the object in front of you is red and another, equally reliable source, tells you that this object is blue. This context's hard information $\mathcal{W}$ contains the propositions $Testimony(R)$ and $Testimony(B)$. And the set of rules $\mathcal{R}$ of $c_7$ includes $r_7 = \dfrac{Testimony(R)}{Believe(R)}$ and $r_8 = \dfrac{Testimony(B)}{Believe(B)}$ with $r_7 \leq r_8$ and $r_8 \leq r_7$, or $r_7 \sim r_8$. Notice that $r_7 \sim r_8$, in effect, says that $r_7$ and $r_8$ have the same weight.

The next step is to specify how the $\bigcirc$-formulas are to be generated from $c_7$ and other weighted contexts. When dealing with regular contexts from Section 1.2, we acquired such formulas from the rules that were *triggered*. Now we add an extra condition: A rule must not only be triggered, but also there must not be another rule that would *outweigh* it. What are the conditions under which one rule outweighs another? Well, the first one is that the two rules must support *contrary* conclusions. And here we need to extend our notion of contrary rules so that it applies to contributory rules too: We'll say that two rules $r = \dfrac{X}{Believe(Y)}$ and $r' = \dfrac{Z}{Believe(W)}$ of some weighted context $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ are contrary in it, written as $contrary_c(r, r')$, if and only if $Y$ and $W$ are inconsistent. The second condition

[24] Just as we did in the case of hedged contexts, we build it into the definition of a weighted context that it doesn't contain multiple rules having the same premise and conclusion, or require that, for any two $r, r' \in \mathcal{R}$, in case $Premise[r] = Premise[r']$ and $Conclusion[r] = Conclusion[r']$, then $r = r'$.

for one rule outweighing another is that both are triggered. And the third that the outweighing rule has at least much weight as the rule that gets outweighed. Once this is realized, the following definition should look very natural.

**Definition 1.9 (Binding rules)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a weighted context. The rules from $\mathcal{R}$ that are* binding *in $c$ are those that belong to the set*

$$\begin{aligned} Binding(c) \;=\; \{r \in \mathcal{R} :\;& r \in Triggered(c) \text{ and} \\ & \text{there is no } r' \in Triggered(c) \text{ such that} \\ & \quad (1)\ r \leq r' \text{ and} \\ & \quad (2)\ contrary_c(r, r')\}. \end{aligned}$$

Let's see this definition at work, when applied to $c_7$. Both rules $r_7$ and $r_8$ are triggered in this context, but neither one of them qualifies as binding. Why doesn't $r_7$ qualify? Well, the rule $r_8$ is as strong as $r_7$ and also contrary to it, $contrary_c(r_7, r_8)$, since $R$ and $B$ are materially inconsistent. And similar reasoning applies to $r_8$. The $\bigcirc$-formulas following from weighted contexts, in turn, will be determined by the following definition:

**Definition 1.10 (Consequence, weighted)** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a weighted epistemic context. Then the statement*

$\bigcirc X$ *follows from $c$ just in case $X \in Conclusion[Binding(c)]$.*

Notice the two differences from the first-pass definition from Section 1.2. First, $Binding(c)$ has been substituted for $Triggered(c)$. And second, $X$ has been changed for $\bigcirc X$ in the final expression, reflecting the change in the structure of rules.

Since the set $Conclusion[Binding(c_7)]$ is empty, neither $\bigcirc Believe(R)$, nor $\bigcirc Believer(B)$ follows from the weighted context $c_7$, just like they didn't follow from the corresponding hedged context $c_3$.

Now let's try to capture the subsequent unfolding of the story. The second episode—that is, the one where you look at the object and see that it's red—can be encoded in the weighted context $c_8 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, where $\mathcal{W}$ contains $Testimony(R)$, $Testimony(B)$, and $Perceive(R)$; the set of rules contains the above $r_7$ and $r_8$, as well as $r_9 = \dfrac{Perceive(R)}{Believe(R)}$; and we have $r_7 \sim r_8 < r_9$. The third and final episode, in turn—the one in which the Epistemology Oracle tells you that the object is blue—can be encoded in the weighted context $c_9 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, where $\mathcal{W}$ is like in $c_8$, except for it also includes $Testimony_O(B)$; $\mathcal{R}$ is like in $c_8$, except for it also includes $r_{10} = \dfrac{Testimony_O(B)}{Believe(B)}$; and we have $r_7 \sim r_8 < r_9 < r_{10}$. It's straightforward to verify—which we won't do here—that the context $c_8$ entails $\bigcirc Believe(R)$ and that the context $c_9$ entails $\bigcirc Believe(B)$. Note that the situation wouldn't be any different if the rules $r_9$ and $r_{10}$ would be present in the first two contexts. That is, were $c_7$ and $c_8$ to have the same set of rules (plus the ordering) as $c_9$ does, the result would still be that, in the first episode, you ought to suspend belief about the color of the object, that, in the second episode, you ought to believe that the object is red, and that, in the third, you ought to believe that the object is blue.

Notice that the two alternative ways of capturing the little epistemic vignette—using hedged contexts and weighted contexts—lead to the same recommendations: neither the formula $\bigcirc Believe(R)$, nor $\bigcirc Believe(B)$ follows from both $c_3$ and $c_7$, only $\bigcirc Believe(R)$ follows from both $c_4$ and $c_8$, and, then, only $\bigcirc Believe(B)$ fol-

lows from both $c_5$ and $c_9$. This prompts the question of whether or not such close correspondence will be observed more generally. We will soon see that the answer to this question is affirmative—with some qualifications—which, in turn, suggests that the views are *extensionally equivalent*. What's more the connections between the views actually run deeper than mere extensional equivalence. But to see this, we must look at the most natural way of understanding reasons and defeat in the model of the contributory-rules view.

### 1.4.2 Reasons, reason strength, and defeat

Again, we take the rules as primitive, understanding them as standing for the favoring relations, and define the other notions in their terms. Suppose we have an agent finding herself in some epistemic situation that we capture using the weighted context $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$. How are to think of the epistemic reasons that the agent has in this situation? Well, here my proposal parallels the one we discussed when thinking about the hedged-rules view: Reasons should be identified with the premises of those rules that are triggered in this context. More specifically, the proposition $X$ is to be defined as a reason for $Y$ in the context $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ if and only if there is a rule $r = \frac{X}{Y}$ in $\mathcal{R}$ that is triggered in $c$. In this case, we would again say that $r$ *provides* a reason for $Y$, that $X$'s being a reason for $Y$ *depends* on $r$, and that $X$ and $Y$ stand in the *reason relation*. To see the proposal at work, consider the weighted context $c_9 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, which expresses the final episode of our running example. We discussed four rules from $\mathcal{R}$, namely, $r_7 = \frac{Testimony(R)}{Believe(R)}$, $r_8 = \frac{Testimony(B)}{Believe(B)}$

, $r_9 = \dfrac{Perceive(R)}{Believe(R)}$, and $r_{10} = \dfrac{Testimony_O(B)}{Believe(B)}$, and noted that all of them are triggered in the context. As a consequence, the premises of all of these rules qualify as reasons that you have. And, to be more precise, $Testimony(R)$ and $Perceive(R)$ are your reasons for believing that the object is red and $Testimony(B)$ and $Testimony_O(B)$ are your reasons to believe that the object is blue.

Contrary to our model of the hedged-rules view, in the contributory rules one it is very natural to talk about various relations that can obtain between reasons. For instance, we can give a precise characterization of the intuitive ideas of when two reasons conflict, as well as when one reason is stronger, or has more weight, than another. Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a context and let $X$ and $Y$ be epistemic reasons in this context, $X$ a reason for $Z$ and $Y$ a reason for $W$. Then $X$ and $Y$ *conflict*, or are *conflicting reasons*, if and only if the rules that they depend on are contrary. What about the relative strength of reasons? Using the same abstract example, we would say that $X$ is *at least as strong* of a reason for $Z$ as $Y$ for $W$ if and only if the rule $r$ that $X$ depends on is at least as strong as the rule $r'$ that $Y$ depends on, that is, if and only if $r' \leq r$. Extension to other relations are straightforward: $X$ as a reason for $Z$ is *strictly stronger* than $Y$ is for $W$ if and only if $X$ is at least as strong as $Y$ and $Y$ is not at least as strong as $X$. $X$ is a weaker reason for $Z$ than $Y$ is for $W$ if and only if $Y$ is stronger than $X$. And the strengths of $X$ as a reason for $Z$ and $Y$ as a reason for $W$ are incomparable if and only if neither $X$ is stronger than $Y$, nor $Y$ is stronger than $X$.

Now let's turn to the notions of defeat and defeaters. We can characterize them

in two equivalent ways. The first appeals to the formal notion of binding rules from previous section; the second proceeds in terms of conflicts between reasons and their relative strengths. Yet again, we fix a weighted context $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ and suppose that, in it, some proposition $X$ is a reason for $Y$ and that $X$ being a reason for $Y$ depends on the rule $r = \frac{X}{Y}$. Then we say that $X$ is *defeated* as a reason for $Y$ in the context of $c$ when the rule $r$ is not among the set of rules that are binding in $c$. Looking at our definition of bindingness, we see that this implies that there is another rule $r'$ in $\mathcal{R}$ that is, first, contrary to $r$, second, triggered in $c$, and, third, at least as strong as $r$ (that is, $r \leq r'$). In this case, we say that the premise of this further rule $r'$ is a *defeater* of $r$ and, hence, $X$ as a reason for $Y$. Alternatively, we can say that $X$ is defeated as a reason for $Y$ in the context of $c$ when there's also a reason $Z$ (for some $W$) in $c$ that's in conflict with $X$ as a reason for $Y$ and that is at least as strong a reason for $W$ as $X$ is for $Y$. In this case, again, $Z$ is a defeater of $X$ as a reason for $Y$.[25]

Consider $c_9$ for illustration, zooming in on the rule $r_8 = \frac{Testimony(B)}{Believe(B)}$. Since it is triggered in $c_9$, the proposition $Testimony(B)$ qualifies as a reason to believe that $B$. But given the fact that $r_8$ is not among the rules that are binding in $c_9$, it is defeated. And just as it was in the case of the corresponding hedged context $c_5$, there are two defeaters. First, the rule $r_7 = \frac{Testimony(R)}{Believe(R)}$ is triggered in the context $c_9$, contrary to $r_8$, and as strong as $r_8$ is, and so its premise $Testimony(R)$ qualifies as a defeater of $r_8$. Similarly, the rule $r_9 = \frac{Perceive(R)}{Believe(R)}$ is triggered, contrary to, and stronger than $r_8$, and so its premise $Perceive(R)$ too qualifies as a

[25] Cf. (Horty 2012, pp. 72–5).

defeater of $r_8$.

# Chapter 2: Connections and implications

## 2.1 Correspondence between the views and some implications

### 2.1.1 Contributory equals restricted hedged

Now we have two models—the model of the hedged-rules view from Section 1.3 and the model of the contributory-rules view from Section 1.4. They offer fairly different pictures of simple epistemic rules, defeat of such rules, as well as cases of epistemic conflict. According to the first model, epistemic rules have built-in hedges that specify the conditions under which the rule doesn't apply. Defeat is understood in terms of these conditions: When one of them obtains, the rule gets defeated and fails to apply. Consequently, any situation where two epistemic rules support contrary conclusions should be thought of as a conflict that's only apparent, because, in it, at most one of these rules applies. According to the second model, epistemic rules are contributory, and they specify what counts in favor of or against the agent having an attitude, as opposed to specifying (individually and independently of other rules present) what beliefs the agent ought to have. Defeat is understood in terms of presence of other rules that support contrary conclusions and are stronger than the given one. Consequently, any situation in which two epistemic rules support contrary conclusions should be thought of as a situation of genuine conflict, but a conflict that is resolvable due to the contributory character of the rules involved.

In spite of these differences, the two models turn out to be equivalent in the

following sense. There's a simple procedure that lets us transform any weighted context into a closely corresponding hedged context (of a particular shape): A consideration $X$ qualifies as a reason in the hedged context if and only if it qualifies as a reason in the original weighted context; $X$ comes out as a defeated reason in the new context if and only if it comes out defeated in the original context; and $\bigcirc X$ follows from the hedged context if and only if $\bigcirc X$ follows from the weighted context. Similarly, there's a simple procedure that lets us transform any hedged context (from a specific class of such contexts) into a corresponding weighted context: Yet again, $X$ is a reason in one context if and only if it is a reason in the other; $X$ is a defeated reason in one context if and only if it is a defeated reason in the other; and $\bigcirc X$ follows from one context if and only if it follows from the other. The equivalence holds under certain natural restrictions, which I want to be explicit about.

Let's start with the contributory-rules view. Our regimented version of the view should be thought of as consisting of two parts. The first part concerns the logic of interaction between contributory rules, and what it says is that it just is the *defeasible* logic we specified in Section 1.4, the logic that lets us get to $\bigcirc$-formulas from any given weighted context. The view's second part concerns the shape of the overall set of weighted contexts, and we can express it as a constraint on an arbitrary weighted context $c$, referring to the contexts that satisfy the constraint as *regular*.

**Definition 2.1 (Regular weighted contexts)** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a weighted context. We say that $c$ is* regular *if and only if, for any two rules $r$, $r' \in \mathcal{R}$ such that $contrary_c(r, r')$, either $r \leq r'$ or $r' \leq r$ (or both).*

This constraint can be seen as expressing the original motivation behind the contributory-rules view, namely, to avoid the counterintuitive conclusion that cases in which simple rules conflict are dilemmic. Since contributory rules were proposed in response to the problem of epistemic conflicts, the constraint seems to be well in place.[1]

Similarly, our regimented version of the hedged-rules view is best thought of as a two-part view. The first part, again, concerns the logic of interaction between hedged rules. And the second part, again, concerns the overall set of hedged contexts. This time, however, there's not one but two constraints a context has to satisfy to qualify as *regular*.

**Definition 2.2 (Regular hedged contexts)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a hedged context. We say that $c$ is* regular *if and only if*

*(1) for any two rules $r, r' \in \mathcal{R}$ such that $contrary_c(r, r')$, either $\neg Premise[r'] \in Hedge[r]$ or $\neg Premise[r] \in Hedge[r']$ (or both); and*

*(2) for any rule $r$ in $\mathcal{R}$, the hedge of $r$ is the set $\{\neg Premise[r'] : r' \in \mathcal{R}'\}$ where $\mathcal{R}' \subseteq \{r' \in \mathcal{R} : contrary_c(r, r')\}$, or, in English, each rule's hedge can contain only negations of the premises of those rules that are contrary to it.*

The first constraint requires little argument. It is, again, just what motivated the hedged-rules strategy in the first place. What it effectively does is ensure that, in any context where two contrary rules get triggered, at least one fail to apply.

[1] Admittedly, the constraint rules out the possibility of epistemic dilemmas for contributory-rules view. Is this problematic? Well, given that the contributory- and the hedged-rules views are both thought of ways of avoiding a commitment to the existence of dilemmas, that a similar constraint will rule out dilemmas for hedged-rules view, and that, in this part of the dissertation, we're mainly interested in the connections between the two views on rules, there doesn't seem to be anything problematic.

What about the second constraint? On the intuitive level, it may be best thought of as a restriction on what rule hedges can do. In effect, it allows that hedges do one thing only, namely, provide a way out of clashes between simple rules. Admittedly, one might argue that this is controversial, but there are some good reasons to have it in place—reasons, that is, that have nothing to do with the fact that we need it to establish the correspondence between the models—and we will return to the issue in Section 2.2. For now, note the following: When ethicists talk about hedged rules—or views resembling the hedged-rules view—they often suggest that rule hedges will only refer to other rules, and the constraint can be naturally seen as a way of capturing this idea in our framework.[2]

To emphasize the closeness of any two corresponding pairs of contexts in the actual statement of the result, it will be useful to introduce the notion of a rule's *counterpart* in a different context. Thus, given a rule $r$ from some context $c$, whether hedged or contributory, we will say that its counterpart in a different context is the rule (if any) that has the same premise as $r$ and the conclusion of which corresponds to that of $r$.

**Definition 2.3 (Counterparts of rules)** *Let $r$ be a hedged rule of the form $\frac{X\ :\ Z}{\bigcirc Y}$ and $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ a weighted context. If there's a rule $r' \in \mathcal{R}$ with $Premise[r'] = X$ and $Conclusion[r'] = Y$, we say that $r'$ is the (weighted)* counterpart *of $r$ in the context $c$, written as $counterpart_c(r) = r'$.*

*Let $r$ be a rule of the form $\frac{X}{Y}$ and $c = \langle \mathcal{W}, \mathcal{R} \rangle$ a hedged context. If there's a*

[2]See, for instance, (Dancy 2004) and (Holton 2002).

*rule* $r' \in \mathcal{R}$ *such that* $Premise[r'] = X$ *and* $Conclusion[r'] = \bigcirc Y$*, we say that* $r'$ *is the (hedged)* counterpart *of* $r$ *in the context* $c$*, written as* $counterpart_c(r) = r'$*.*

The result itself are the following two observation—both of which are verified in the Appendix:

**Observation 1** *Let* $c = \langle \mathcal{W}, \mathcal{R} \rangle$ *be a regular hedged context. Then there is a regular weighted context* $c' = \langle \mathcal{W}', \mathcal{R}', \leq \rangle$ *such that*

*(1)* $\mathcal{W}' = \mathcal{W}$*;*

*(2) for every rule* $r' \in \mathcal{R}'$*, there's a counterpart rule* $r \in \mathcal{R}$*;*

*(3)* $X$ *is a reason for* $Y$ *in* $c$ *if and only if* $X$ *is a reason for* $Y$ *in* $c'$*;*

*(4)* $X$ *is defeated as a reason for* $Y$ *by a consideration* $Z$ *in* $c$ *if and only if* $X$ *is defeated as a reason for* $Y$ *by* $Z$ *in* $c'$*; and*

*(5)* $\bigcirc X$ *follows from* $c$ *if and only if* $\bigcirc X$ *follows from* $c'$*.*

**Observation 2** *Let* $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ *be a regular weighted context. Then there is a regular hedged context* $c' = \langle \mathcal{W}', \mathcal{R}' \rangle$ *such that*

*(1)* $\mathcal{W}' = \mathcal{W}$*;*

*(2) for every rule* $r' \in \mathcal{R}'$*, there's a counterpart rule* $r \in \mathcal{R}$*;*

*(3)* $X$ *is a reason for* $Y$ *in* $c$ *if and only if* $X$ *is a reason for* $Y$ *in* $c'$*;*

*(4)* $X$ *is defeated as a reason for* $Y$ *by a consideration* $Z$ *in* $c$ *if and only if* $X$ *is defeated as a reason for* $Y$ *by* $Z$ *in* $c'$*; and*

*(5) $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.*

Observations 1 and 2 reveal a very close connection between the two models. But what do they tell us about the relations between the views that these two models represent? For starters, if the two models represent the views adequately—as I think they do—then the observations, through points (5), certainly show that the views are extensionally equivalent, or that the views captured by the models provide the same answers to the question of what beliefs the agent ought to have in some given epistemic situation. While the extensional equivalence of the views might not come as that much of a surprise—in the end, they were proposed as responses to the same problem—the observations establish more than a mere extensional equivalence . First, points (3) and (4) suggest that there's a one-to-one correspondence between other normative notions in these views. If we look back at how the concepts of a reason and defeat were defined in Sections 1.3.3 and 1.4.2, we should find this quite surprising. While the definitions of the first notion, reason, run parallel, the same can't be said about the notion of defeat. And second, the two observations also reveal the *conditions* under which the hedged-rules view and the contributory-rules view are extensionally equivalent: The correspondence holds only for those contexts that satisfy the constraints expressed in the definitions of *regular* contexts. One constraint (the no-dilemmas constraint) is shared by both types of contexts, but the other one isn't. So it reveals what shape rule hedges must have for the hedged-rules view and the contributory-rules view to be equivalent. Also, note the following: With this constraint in place, we are dealing with a restricted version of the hedged-rules

view; and it is the restricted version that's equivalent to the contributory-rules view. We will discuss the full version of the view in Section 2.2 and see that it is more expressive.

Before we do that, however, I want to show how the close connections between the views revealed by the result can be used to show that some received ideas about the hedged-rules view are mistaken. So, in the remainder of this section, I will zoom in on *two* phenomena that the hedged-rules view is standardly taken to fail to account for; explain how they can be accounted for in the contributory-rules view; and then mimic the explanation in the hedged-rules view. What are the two phenomena? Well, the first one is more transparent in the moral domain. It's what we might call the *residual badness* that's observable in at least some situations where the agent acts just as she ought and yet still has some reason to regret not having acted otherwise. The second phenomenon is what we might call the *composition of reasons*: It'd seem that there are at least some situations where two (or more) weaker considerations *together* defeat a consideration that's stronger than each of them taken in isolation.[3]

### 2.1.2 Residual badness and regret

While the importance of residual badness and the regret that comes with it is widely recognized in moral philosophy, the status of this phenomenon in epistemology is less clear. In fact, the question of whether or not it even needs to be

[3]Both residual badness and composition of reasons are often brought up in arguments against the hedged-rules view, but, as we will see, wrongly so. Dancy (2004), for instance, associates the hedged-rules view with Holton (2002) and Scanlon (1998, 2000) and points to the two phenomena to argue against it—see especially (Dancy 2004, pp. 22–9).

accounted for in epistemic contexts is controversial. I suspect that, upon reflection, most epistemologists would say that the phenomenon is real, even if they would add that the epistemic situations that involve residual bad-making features are less common and far more unusual than the situations that involve such features in the moral domain. In the end, the recent work of Christensen (2007a, 2010a, 2013) on higher-order evidence and related issues seems to have convinced most people that there *are* peculiar epistemic situations such that, even if the agent responds to the them by adopting the doxastic attitudes that she ought, there's still something bad or regrettable about the doxastic state that she ends up in.[4] But not all epistemologists are convinced. Bradley (2019), for instance, has suggested that if the agent does what she ought to (epistemically) and doesn't comply with a *defeated* epistemic rule as a result, the situation doesn't involve any residual bad-making features and there's no place for regret, whether this regret be epistemic or otherwise.[5] I myself side with the majority (or what I think the majority is), but nothing in my argument will depend on this. For what I will do is establish the following conditional claim: If there are epistemic situations involving residual bad-making features, then we can account for them using contributory rules, as well as hedged rules.

Given that the phenomenon is less controversial in ethics, it will be more

[4]Christensen's own stance on the issue appears very clearly in the following passage: "[..I]t seems to me that we should continue to recognize a sense in which there is often something epistemically wrong with the agent's beliefs after she takes correct account of [higher-order evidence]. There's something epistemically regrettable about the agent's being prevented, due to her giving [higher-order evidence] its proper respect, from following simple logic, or from believing in the hypothesis that's far and away the best explanation for her evidence" (Christensen 2010a, p. 204).

[5]See (Bradley 2019, p. 9). I don't find Bradley's argument at all convincing. All he does is appeal to our intuition in situations where the defeated rule is strongly undermined—we'll take a closer look at such cases in Section 2.2.1. But, clearly, the fact that it intuitively seems to us that there's no place for regret in this type of situations doesn't establish that there there's no place for regret in any type of situation.

convenient for us to shift attention from epistemic contexts to moral ones, as well as from simple epistemic rules to their counterpart in the moral domain, what we might call *moral duties* or *moral principles*. Not surprisingly, this means that we need to make some slight adjustments to the interpretation of the formal language we have been using. Most importantly, we must broaden the meaning of the deontic operator $\bigcirc(\cdot)$. Henceforth, a formula of the form $\bigcirc X$, with $X \neq \mathit{Believe}(Y)$ for some $Y$, should be read thus: According to the dictates of morality, it ought to be the case that $X$, or, simply, it morally ought to be the case that $X$. Above we used the notion of a context to express epistemic situations, but we can just as well use it to capture moral situations. One important difference between epistemic and moral contexts is that the conclusions of the rules of the latter won't include any formulas of the form $\mathit{Believe}(X)$. For moral duties or principles will speak to what ought to happen, as opposed to what beliefs the agent ought to have.

Now let's zoom in on the phenomenon, starting with the following timeworn example:

> **Drowning Child**: You have promised your friend Taylor to have a dinner with her. Your route to the restaurant takes you past a pond, and, as you are walking past it, you notice that a child has fallen in. The child is crying in distress, and all your evidence suggests that it is going to drown, unless you do something about it. However, if you rescue the child, you will get your clothes wet and muddy, and won't make it to the dinner with Taylor.

Perhaps everyone would agree that you ought to save the child, thereby missing the dinner with Taylor. So were you to save the child, you'd be acting as you ought. However, even if you did that, you'd likely still feel some regret about letting Taylor down by not keeping your promise to her. In fact, it seems that such a feeling of regret would be appropriate, and that you would owe her an explanation and an apology.

It's widely agreed that a major selling point of contributory moral principles, what W. D. Ross (1930) called *prima facie duties*, is that they naturally account for residual bad-making features and regret involved in situations like the Drowning Child. How? Well, let's say that you think of the ethical system as consisting of duties, involving the duty to keep one's promises and the duty to help those who are in need. Then we would say that, in the particular circumstances that obtain in Drowning Child, the duty to save the child is much stronger, and that it outweighs the duty to keep promises. However, the latter duty remains in place, in spite of being outweighed, and the presence of this residual duty explains why the feeling of regret and an apology to Taylor are appropriate.

Let's now see how this can be captured precisely in our model of the contributory-rules view, or, rather, the view on which moral principles are contributory. The first step is to express the case in a (moral) context. Let *Promise* stand for the proposition that you have promised to dine with Taylor; *Dine* for the proposition that you dine with her; *Drowning* for the proposition that the child is drowning in the pond and needs your help, and *Save* for the proposition that you save the child. We assume that *Save* and *Dine* are materially inconsistent. The relevant instance of the

duty to keep one's promises can be captured by means of the rule $r_{11} = \frac{Promise}{Dine}$, saying that your promise to Taylor is a reason for you to dine with her, or that, by default, you ought to dine with Taylor if you have promised to do so. The relevant instance of the duty to help those in need, in turn, can be captured by means of the rule $r_{12} = \frac{Drowning}{Save}$, saying that the child's drowning is a reason for you to save it. The entire situation can then be encoded in the context $c_{10} = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, where the hard information $\mathcal{W}$ is comprised of $Promise$ and $Drowning$, while $\mathcal{R}$ includes the rules $r_{11}$ and $r_{12}$ with $r_{11} < r_{12}$. It's easy to check that the formula $\bigcirc Save$ follows from $c_{10}$, supporting the intuitive conclusion that you ought to save the child.

But we may also want to capture the residual badness involved in this situation and others like it. One straightforward way to do so is by associating it with what we might call *residual reasons*, or reasons that are in force in the situation and yet do not have a corresponding ought, or reasons that get outweighed by stronger ones. Accordingly, we can say that, given some weighted context $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, regret about not $Y$-ing is appropriate if, in $c$, there's at least some reason $X$ to $Y$, and yet all the reasons for $Y$-ing are defeated. The fact that $r_{11} = \frac{Promise}{Dine}$ is triggered in $c_{10}$ means that you have a reason to dine with Taylor, namely, $Promise$. But this reason is defeated by a stronger one, namely, $Drowning$, and, given that you don't have other reasons to dine with Taylor, all of your reasons to dine with her are defeated. And that's enough to conclude that regret about not-$Dine$ is appropriate.

Can we capture residual badness and regret in the hedged-rules model? It turns out that we can, and quite easily. First, let's express the Drowning Child in a hedged

context. Take the context $c_{11} = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ is the set $\{Promise, Drowning\}$ and $\mathcal{R}$ includes the rules

$$r_{13} = \frac{Promise : \neg Drowning}{\bigcirc Dine}, \text{ as well as}$$

$$r_{14} = \frac{Drowning}{\bigcirc Save}.$$

It's easy to see that the formula $\bigcirc Save$ follows from $c_{11}$, and that $Promise$ comes out as a reason for $Dine$, as the only one and also as a defeated one. And there appears to be nothing in the way of reusing the above statement of the conditions for when regret is fitting: Given some regular hedged context $c = \langle \mathcal{W}, \mathcal{R} \rangle$, regret about not $Y$-ing is appropriate if there's at least some reason to $Y$, and yet all the reasons for $Y$-ing are defeated. With this, the regret about not dining with Taylor is, again, well in place.

So it looks like we can capture residual badness and regret not only in the contributory-rules model, but also in the hedged-rules one. This suggests that the received opinion that one has to appeal to contributory rules (or duties) in order to account for regret is mistaken, and that the hedged-rules view is, in fact, sufficiently versatile to account for it too. A question remains, however: What sustains the common opinion about the hedged-rules view and regret? Why hasn't the move we have just made been made before? Well, my best guess is that this is due to the *implicit* and very strong assumptions about the shape of rules (or moral principles) and the logic that governs their interaction. For illustration, let's take a look at the way Richard Holton (2002) introduces moral principles—which exemplifies a very common way of thinking. Holton invites us to imagine a set of descriptive, but

morally relevant one-place predicates $\{F_1(\cdot), F_2(\cdot), \ldots, F_m(\cdot)\}$, such as "is a killing", "is done in self-defense", as well as a moral predicate $F_c(\cdot)$, such as "you (morally) should do" or "you are (morally) permitted to do." He then goes on to suggest that these predicates can occur in a moral principle of the form

$$\forall x([F_1(x)\&F_2(x)\&\cdots\&F_m(x)] \supset F_c(x)),$$

as well as in a corresponding set of descriptive sentences $\{F_1(a), F_2(a), \ldots F_m(a)\}$. Two specific principles he considers run thus:

(1) $\forall x(x$ is a killing $\supset$ you shouldn't do $x)$,

(2) $\forall x(x$ is a killing & $x$ is done is self defence $\supset$ you may do x).

From here, it should be easy to see how Holton thinks of a principle applying (or not applying) in a given situation: A principle applies in a situation if and only if the situation—that's conceived of as a set of descriptive propositions—can be subsumed under the principle—conceived of as a universally quantified proposition.[6] Why is this important? Well, on any way of thinking about principles sufficiently close to this one, a principle will always either fully apply in a situation, or it will wholly fail to apply, and there will be no place for any sort of middle ground in between. And without such middle ground, there's no place for residual duties that could account for residual badness and regret. In our model, by contrast, a rule's premise is distinct from its hedges. As a result, we can distinguish between a rule not applying at all

[6] Admittedly, Holton's own view on principles is more nuanced. He goes on to suggest that each principle, as well as each description of a moral situation must also include a special *That's it* clause. However, this is orthogonal to the point I'm making in the main text.

and a rule not applying because one of the conditions specified in its hedge obtains and associate regret with the latter type of cases.

Notice what we have arrived at: Whether we are dealing with a weighted context or a hedged one, any situation involving multiple triggered rules supporting opposing conclusion will also involve residual bad-making and regret. Admittedly, it is natural to wonder if regret is really appropriate in any situation of conflict between rules or principles, and especially so when conflicts are epistemic. In Section 2.2 we'll consider a situations where regret does *not* seem appropriate and discuss how it fits into the overall picture. Before we do that, however, let's look at the other phenomenon that's standardly taken to cause trouble for the hedged-rules view.

### 2.1.3 Composition of reasons

The vexed topic of composing reasons has been discussed in both ethics and epistemology.[7] However, only in the ethics literature has it been linked to the different views on rules and argued to cause trouble for the hedged-rules view. So we will, again, focus on the practical domain. Consider the following example, adapted from Campbell Brown (2014):

> **Whiskey vs. Water**: Imagine you have to decide whether to drink whiskey or water. Each option is supported by exactly one reason. If you choose whiskey, you'll be happy; if water, healthy. Plausibly, you ought to do whatever is supported by the strongest reason, and let's say this is

[7] For (formal) work on this topic see (Brown 2014), (Delgrande & Schaub 2004), (Gómez Lucero et al. 2009, 2013), (Horty 2012, 91–5), (Nair 2016), (Pollock 1995, pp. 101–2), (Prakken 2005, 2019), (Schroeder 2007, pp. 123–45).

> drinking whiskey (happiness outweighs health). But now suppose there's a second reason for drinking water: it's cheap. This new reason, like the old one, is less strong than the reason for drinking whiskey (happiness outweighs money). So, as before, drinking whiskey is favored by the strongest reason. Yet this may no longer be what you ought to do. By combining together considerations of health and money would seem to defeat the stronger consideration of happiness that neither could defeat alone.[8]

Now, how are we to make sense of the idea that considerations of health and money combine to give you sufficient reason to drink water? The proponents of contributory rules are standardly taken to have an answer ready at hand: Supposing, purely for the sake of simplicity, that there are simple rules corresponding to all the considerations at play in Whiskey vs. Water and that these rules are contributory, one can say that the weights of the pro-water rules combine to trump the pro-whiskey rule. The proponents of hedged rules, on the other hand, are supposed to have a hard time answering this question: Supposing that the rules at play in the situation are all hedged, it is hard to resist the following conclusion. Were *only* the considerations of health and happiness to be relevant, we would surely say—or would have to say—that the situation at hand is one where a condition specified in the hedge of the pro-water rule obtains, and that this rule fails to apply as a result. What's more, we'd say the exact same thing about the other pro-water rule in a situation where *only* the considerations of money and happiness are relevant. But, then, it's

[8]See (Brown 2014, pp. 779–80).

hard to resist the conclusion that the overall situation of Whiskey vs. Water is one where the pro-water rules fail to apply, and that, therefore, there's no way they could defeat the pro-whiskey rule.

This line of reasoning appears to be quite convincing at first, but the proponents of the hedged-rules view do actually have a way to respond to it. This becomes manifest once we look at the case at hand through the lens of our formal framework. The first step will be to see how reason composition can be captured in the model of the contributory-rules view, the second to extrapolate it to the model of the hedged-rules view.

Let $Whiskey$ stand for the proposition that you drink whiskey; $Water$ for the proposition that you drink water; $Happiness$ for the proposition that drinking whiskey would make you happy; $Health$ for the proposition drinking water would make you healthy; and, finally, $Money$ for the proposition that, were you to drink water, you would save money. We will also assume that $Whiskey$ and $Water$ are materially inconsistent.[9] Now we could try encoding Whiskey vs. Water in the weighted context $c_{12}$ where $\mathcal{W} = \{Happiness, Health, Money\}$ and $\mathcal{R}$ includes the rules $r_{15} = \frac{Health}{Water}$, $r_{16} = \frac{Money}{Water}$, as well as $r_{17} = \frac{Happiness}{Whiskey}$ with $r_{15} < r_{17}$ and $r_{16} < r_{17}$. What do the rules say? Well, the first one says that the fact that water would make you healthy speaks in favor of drinking water; the second says that the fact that water would let you save money speaks in favor of drinking water; and, finally, the third rule says that the fact that whiskey would make you happy speaks

[9] We can easily fill in the details of the story in a way that would justify the constraint that you can't drink water as well as whiskey. Perhaps, water is relatively expensive and you simply don't have enough money to afford both.

in favor of drinking whiskey. It's not difficult to see that the context $c_{12}$ entails the formula $\bigcirc Whiskey$ supporting the conclusion that you ought to drink whiskey. But this is, of course, the wrong result. For we stipulated that, in the case at hand, what you ought to do all things considered is drink water.

But it shouldn't be all that surprising that $c_{12}$ delivers the wrong result, as the information that considerations of health and money combine is utterly absent from it. I know of only one way to supplement the context with it, and it's due to Horty (2012) who has used a framework that's similar to ours to model interaction between reasons. Horty considers the question of composing reasons only briefly, suggesting that we model it by supplementing contexts with additional rules.[10] Adjusting the idea to our setting, this would amount to supplementing $c_8$ with the pro-water rule $r_{18} = \dfrac{Health\&Money}{Water}$ and setting it to be stronger than the pro-whiskey rule $r_{17}$. The result is the contexts $c_9 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ where $\mathcal{W}$ is just as it was before, $\mathcal{R}$ now also includes $r_{18}$, and the priority relation tells us what it did before, as well as that $r_{17} < r_{18}$. Unlike $c_{12}$, the updated context $c_{13}$ does entail the formula $\bigcirc Water$, supporting the intended conclusion that you ought to drink water.

There's reason not to be entirely satisfied with this way of handling the case: Nothing of importance would seem to hinge on us supposing that $r_{15}$, $r_{16}$, and $r_{17}$ are instances of some simple moral principles, *prima facie* duties, or whatever we think is the analogue of simple epistemic rules in the moral domain. But the rule $r_{18}$ can't be interpreted in this way. Instead, we must think of it as being derived from the simple $r_{15}$ and $r_{16}$, and derived by a mechanism that lies outside

[10] See the discussion in (Horty 2012, pp. 91–4).

our formal model.[11] The defeasible logic framework we have been using has plenty of benefits—in the end, it let us formulate two precise, concrete, and transparent theories about the shape of simple rules and the way such rules interact in situations of conflict—but it simply doesn't have much to say about the way simple rules combine. The main reason is that the phenomenon itself is highly irregular. In some cases—like Whiskey vs. Water—two reasons supporting the same conclusion will indeed combine, but in other cases two reasons supporting the same conclusion might not only fail to combine, but may actually diminish each other's initial force, or cancel it entirely. The formal framework is good for capturing regularities, but if there are no regularities in the phenomenon, then it shouldn't be all that surprising that the framework doesn't provide insights into it. Nevertheless, the phenomenon itself is real, and, while not having a worked-out theory about it, we can still capture it in part by including the rule $r_{18}$ in the context and suggesting that it is derived from the two pro-water rules $r_{15}$ and $r_{16}$.

Now, in light of the result established in Section 2.1, we know that there is a hedged context corresponding to $c_{13}$ that contains a counterpart for each rule in $\mathcal{R}$ and delivers the exact same result. What does it look like? Well, it's the context $c_{14} = \langle \mathcal{W}, \mathcal{R} \rangle$, where $\mathcal{W}$ is just what it was in the case of $c_{13}$, and where $\mathcal{R}$ includes the following rules:

$$r_{19} = \frac{Health\colon \neg Hapiness}{\bigcirc Water},$$

$$r_{20} = \frac{Money : \neg Hapiness}{\bigcirc Water},$$

[11] Cf. (Horty 2012, pp. 93–4).

$$r_{21} = \frac{Hapiness : \neg(Health\&Money)}{\bigcirc Whiskey}, \text{ and}$$

$$r_{22} = \frac{Health\&Money}{\bigcirc Water}.$$

Of course, the formula $\bigcirc Water$ follows from $c_{14}$, suggesting that the hedged-rules view can, in fact, capture cases where weaker considerations combine to defeat a stronger one, much like the contributory-rules view can. The rule $r_{22}$, again, shouldn't be thought of as an instance of some simple rule schema, but as one that's derived from the simple rules $r_{19}$ and $r_{20}$. What's more, the context $c_9$ would seem to put us in a position to identify the crucial assumption in the reasoning to the conclusion that the hedged-rules view can't account for composition of reasons: Zoom in on some hedged rule $r$ and take any situation satisfying at least one of the conditions specified in the hedge of $r$. Then, it's not only that the conclusion of $r$ won't get detached in this situation, but also that $r$ can't have *any other* effects on the situation. In particular, $r$ can't contribute to the derivation of another stronger rule with the same conclusion. The preceding discussion suggests that the proponent of the hedged-rules view can deny this assumption and say that what happens in cases like Whiskey vs. Water is that the two pro-water rules combine to create a stronger rule, in spite of the fact that they fail to apply, in the sense that their conclusions don't get detached.

In case you are dissatisfied with this way of handing cases involving composing reasons, I have one more consideration for you. I think that our discussion makes the following clear: If the phenomenon of composing reasons poses a problem at all, then it is a problem for both the hedged-rules model and the contributory-rules one.

This, in turn, suggests that with regard to this phenomenon—just as it was with regard to the phenomenon of residual badness and regret—the hedged-rules view does as well as the contributory-rules view.

## 2.2 Undermining

Hitherto all the cases of epistemic (and nonepistemic) conflicts we have considered were cases where simple rules supported contrary conclusions. But cases of this sort do not exhaust the space of epistemic conflicts. For there are also situations where simple rules *get undermined*. Just suppose that you find yourself in a situation where you are looking at what seems to be a red object and where the Epistemology Oracle tells you that Perception—Hedged Perception, Contributory Perception, or whatever the correct rule turns out to be—is not a genuine rule. It's very natural to think that in this case the Oracle's testimony defeats Perception, that, therefore, it's not the case that you ought to believe that the object in front of you is red, and that this situation is different from the one where the Oracle tells you that the object in front of you is blue. Our treatment of epistemic conflicts would be incomplete if we didn't discuss the way the two different views on rules might handle situations of this sort.

### 2.2.1 The limits of the contributory-rules view

It'll be useful to have a concrete scenario involving undermining defeat on the table. Consider the following case, due to Jonathan Dancy:

> **Drug**: Suppose that you have taken a drug that makes blue things look red and red things look blue to you. Furthermore, you have all the evidence to believe that you have taken this drug and you know how it works. Now you look at an object in front if you. It looks red.[12]

Intuitively, what you ought to do in this situation is believe that the object in front of you is blue. What's more, it also seems that you don't have the slightest reason to believe that this object is red. Normally, an object's looking red would provide reason for believing that it is red, but in this case the fact that you've taken the drug and your knowledge of its workings would seem to reduce the effects that your perception would normally have to nothing.

Dancy (2004, 2017) appeals to cases like Drug—as does Bradley (2019) following him—to argue that views on which rules are contributory can't be correct.[13] The issue is supposed to be that, when we apply some contributory version of Perception to Drug, we end up with the conclusion that you actually do have a reason to believe that the object is red, or that there's at least something speaking in favor of you believing that the object is red, which appears counterintuitive. Note too that the problem is supposed to be fully general: Contributory rules will deliver wrong verdicts in all structurally similar cases, as they "[..] say that a feature that is a reason to believe $X$ is *always* a reason to believe $X$" (Bradley 2019, p. 8, my emphasis). Some authors do not accept this pessimistic verdict.[14] But let's bracket

[12]The case is adapted from (Dancy 2017).

[13]Dancy himself, of course, doesn't talk about contributory rules. He uses the case in his main argument against the view called *generalism*, according to which there are no universal moral principles.

[14]See e.g., (Horty 2012, Ch. 6) and (McKeever & Ridge 2006).

that for a minute, and see what our formal model of the contributory-rules view has to say about the case.

What weighted context would we use to capture the case? Let's start with the hard information. Let $R$ be as before—that is, the proposition that the object in front of you is red—and let $Drug$ stand for the proposition that you've taken a drug that has the effects described in the passage. Then, the context's hard information would include the formulas $Perceive(R)$ and $Drug$. What about the rules? First off, we need an instance of the Perception schema, or the rule $r_9 = \dfrac{Perceive(R)}{Believe(R)}$, which has to be present in any situation where $Perceive(R)$ obtains. But if we don't want the formula $\bigcirc Believe(R)$ to follow from the context, we need an additional rule that would defeat $r_9$. A very natural candidate is the rule $r_{23} = \dfrac{Perceive(R)\&Drug}{Believe(B)}$ which we can make sense of in the way we made sense of derived rules when discussing composition of reasons: In the case at hand, the information about the drug's effects and the fact that the object looks red to you combine, jointly speaking in favor of you believing that the object is blue. Unfortunately, we don't have a good way to model the composition process, and so we resort to a black box standing for it, a black box in the form of $r_{23}$. Putting this together, we acquire the weighted context $c_{14} = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, where $\mathcal{W} = \{Perceive(R),\ Drug\}$ and $\mathcal{R}$ includes $r_9$ and $r_{23}$, with $r_9 < r_{23}$. It's easy to check that the formula $\bigcirc Believe(B)$ follows from $c_{14}$. But, while this is a welcome result, the context $c_{14}$ doesn't actually give us all that's intuitively desirable: Since the rule $r_8$ is triggered in the context, $Perceive(R)$ qualifies as a reason to believe $R$, even if a defeated one. So the model implies that there's at least some reason for you to believe that the object is red. What's more,

our discussion of regret in Section 2.1.2 invites a further counterintuitive conclusion: It is appropriate for you to regret not believing that the object is red. (Generalizing from $c_{14}$, it's difficult to see how we could avoid the counterintuitive conclusions, given how limited the resources of the model of the contributory-rules view are: Being an instance of the Perception schema, $r_8$ has to be in place in any context representing Drug. And given that the object in front of you looks red to you, $r_8$ will get triggered. If it is not to be among the binding rules, the context must also include some rule $r$ that's contrary to $r_8$ and at least as strong as it is. But the claim of $r_8$ will remain standing in the presence of any such rule.)

Now let's switch gears and try to capture Drug, using the hedged-rules view. Consider the hedged context $c_{15}$ where $\mathcal{W}$ contains the formulas $Perceive(R)$ and $Drug$ and where $\mathcal{R}$ includes the rules

$$r_{24} = \frac{Perceive(R)\colon\ \neg Drug}{\bigcirc Believe(R)} \text{ and}$$

$$r_{25} = \frac{Perceive(R)\&Drug}{\bigcirc Believe(B)}\,.$$

Yet again, $r_{24}$ should be thought of as an instance of the Perception schema and $r_{25}$ as a derived rule. It's easy to see that $\bigcirc Believe(B)$ follows from $c_{10}$, as desired. Unfortunately, however, $Perceive(R)$ again comes out as a reason to believe $R$.

Now, my strategy for avoiding the coutnerintuitive conclusion will be to revise the notion of defeat from Section 1.3.3, distinguishing between two types of defeat and associating only one of them with residual reasons (and regret that may or may not come with them). Recall how defeat was defined. Given a context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ and a rule $r$ that's triggered in it, we said that $r$ is defeated by some consideration

$Z$ if and only if $\neg Z \in Hedge[r]$ and $Z$ follows from the context's hard information. Recall too when two rules qualify as contrary. Given a context $c = \langle \mathcal{W}, \mathcal{R} \rangle$, two rules $r = \dfrac{X}{\bigcirc Believe(Y)}$ and $r' = \dfrac{Z}{\bigcirc Believe(W)}$ from $\mathcal{R}$ are said to be contrary just in case $Y$ and $W$ are inconsistent. With these two notions in hand, we can actually capture the distinction between *rebutting* and *undermining* defeat in the model, as follows:

**Definition 2.4 (Rebutting and undermining defeat, hedged view)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a hedged context and $r$ a rule from $\mathcal{R}$ that's triggered in it. Then:*

*(1) $r$ is* rebutted *by a consideration $Z$ in $c$ if and only if $r$ is defeated by $Z$ and $Z = Premise(r')$ for some $r' \in \mathcal{R}$ such that $contrary_c(r, r')$;*

*(2) $r$ is* undermined *by a consideration $Z$ if and only if $r$ is defeated by $Z$ and there is no $r' \in \mathcal{R}$ such that $Z = Premise(r')$ and $contrary_c(r, r')$.*

How does this distinction help with the Drug scenario? Well, I propose that we slightly modify the conditions under which a consideration $X$ qualifies as a reason for $Y$. Henceforth, we will say that $X$ qualifies as a reason for $Y$ in a context $c$ only in case the rule $r$, which $X$'s being a reason for $Y$ depends on, is both triggered and not undermined in the context $c$. (Note that in case the rule $r$ gets rebutted, $X$ will still come out as a reason for $Y$.) It's easy to see that the $r_{24}$ is defeated in the context $c_{15}$. For the proposition $Drug$ qualifies as its defeater: We have both $\neg Drug \in Hedge[r_{24}]$ and $\mathcal{W} \vdash Drug$. However, $Drug$ is not a premise of some other rule that would be contrary to $r_{24}$, and so it is an undermining, and not a rebutting defeater.

At this point you may wonder if a similar sort of move, to refine the notion of defeat, can't be made in the contributory-rules model. Well, it can't, and the result from Section 2.1 can be used to explain why. As a first step, note that the context $c_{15}$ does not actually qualify as a *regular* hedged context. Recall the second constraint on regular contexts: If $c = \langle \mathcal{W}, \mathcal{R} \rangle$ is regular, then the hedge of any rule $r$ from it can contain *only* the negations of the premises of those rules from $\mathcal{R}$ that are contrary to it. But the formula $\neg Drug$ is not the negation of any rule from $\mathcal{R}$, let alone a rule that's contrary to $r_{24}$. So while $c_{15}$ appears to be a perfectly fine representation of the case, it is not regular. Recall that the result from Section 2.1 established that the contributory-rules view is equivalent to a *restricted* version of the hedged-rules view. Drug can be captured using hedged rules, but we have to go beyond the restricted version, or use a context that's not regular, to do so.

The fact that $c_{15}$ doesn't have a weighted counterpart may indicate that we won't be able to adequately capture Drug in the contributory-rules model, but it doesn't firmly establish it. So here's an argument that does: Our result shows that every weighted context has a hedged counterpart that's regular. But the only type of defeat that regular hedged contexts allow for is rebutting defeat. So weighted contexts allow only for rebutting defeat, and one needs undermining defeat in order to capture Drug. The following observation supports the crucial second premise—its simple proof is given in the Appendix:

**Observation 3** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a regular hedged context and $r$ a rule from $\mathcal{R}$ that's triggered in $c$. Then, if $r$ is defeated in $c$, it is rebutted.*

So our formal model supports the claims of Dancy and others that the contributory rules-view misdiagnoses cases like Drug, and that, therefore, the view just can't be correct.

I can think of two ways the proponents of contributory rules might try resisting these claims. The first is to point out the seeming arbitrariness in our definitions of rebutting and undermining defeat: Why did we associate the former with residual reasons (and regret), rather than the other way around? Couldn't we just as well have stipulated that $X$ favors $Y$ only if the rule that $X$'s being a reason for $Y$ depends on is not defeated? The answer to this question is simple: We could have done that, and it would give us the intuitively correct result in Drug, but only at expense of producing counterintuitive results in a whole plethora of other cases, that is, all of the epistemic (and, perhaps, nonepistemic) situations where residual reasons have a place. The second way to resist the claims is by suggesting that our formal model of the view is defective and doesn't capture it in full. This line of thought requires a more complex response.

### 2.2.2 Mixed view

Suppose a proponent of the contributory-rules view insists that our model of the view is defective. What should we say in response? Well, I think we should invite them to explain what the model is lacking and, then, once they do that, we would add the missing piece to the model. But one thing to look out for in this process of enhancing the model is that the model we end up with is still a model

of the contributory-rules view, as opposed to a *mixed view*, on which rules are both contributory and hedged. Let me offer an illustration of how this can happen.

Dancy's Drug scenario is discussed in some detail in (Horty 2012, Sec. 6). Working with a defeasible logic framework that's similar to, but also more expressive than, ours, Horty develops a model that is naturally though of as a model of interaction between contributory rules, much like our model from Section 1.4.[15] And he encodes the Drug case into a contexts that does deliver the intuitive verdict that the right thing for you to do is to believe that the object in front of you is blue, as well as that there's absolutely no reason to believe that it is red.[16] How? Well, the model allows for a special type of *exclusionary rules* which, when triggered, take other rules out of consideration. Horty's representation of the Drug is similar to our $c_{14}$, except for that it also includes an exclusionary rule that's triggered by the proposition $Drug$, and that takes the (counterpart of the) rule $r_9 = \dfrac{Perceive(R)}{Believe(R)}$ out of consideration, nullifying its effects. I don't have anything against Horty's representation of the scenario, but I also think that the view on rules it implies is not the simple contributory one anymore. Why? Well, if we have an exclusionary rule $r$ that gets triggered when $X$ obtains and that takes some other rule $r'$ out of consideration, then we can just as well think of $X$ as a circumstance under which $r'$ fails to apply.[17] Or, to put it in more familiar terms, we can just as well think of $r'$ as a hedged rule with $\neg X$ in the hedge, and forgo even mentioning the spe-

[15]Horty himself presents it as a model of the way *reasons* interact to support conclusions. However, rules are basic in the framework, just as they are in ours.

[16]See, especially, (Horty 2012, pp. 231–2).

[17]I'm glancing over some details here: One exclusionary rule can be taken out of consideration by another one, which means that $r$ being triggered doesn't yet imply that $r'$ will be taken out of consideration. These details are not important for the point I'm making.

cial exclusionary rule $r$. However, in light of the fact that $r'$ is a contributory rule, Horty's model seems to entail a mixed view on rules, according to which rules are *both* contributory and hedged.

Where does this leave us with regard to the idea that we couldn't capture Dancy's case because of the limitations of the formal model, not the (simple) contributory view? I can't claim to have shown that this idea is wrong, but there appears to be very good reason to be skeptical about it. What's more, it'd seem that the proponent of contributory rules herself should be interested in exploring the mixed view—since, at this point, conceding that contributory rules can have hedges seems like a dialectically more promising option than insisting that the formal model is inadequate. So that's what we are going to do in the rest of this section: formulate and explore the mixed view in our framework.

It shouldn't come as surprise that all we need to do to capture this view in our framework is combine ideas from Sections 1.3 and 1.4. First off, the contributory and hedged versions of Testimony and Perception will have the following form:

$$\frac{Testimony(X) : \neg Z_1, \ldots, \neg Z_n}{Believe(X)} \qquad \frac{Perceive(X) : \neg Z_1, \ldots, \neg Z_m}{Believe(X)}$$

What makes them different from the corresponding (absolute) hedged schemas are the conclusions, which are formulas of the form $X$, as opposed to $\bigcirc Believe(X)$. And what makes them different from the corresponding (simple) contributory schemas is the presence of a hedge.

Epistemic situations will be encoded in what we'll call *mixed epistemic contexts*. Any such context is a triple $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ where $\mathcal{W}$, again, states what the

descriptive features of the situation; $\mathcal{R}$ contains instances of rule schemas like the above two; and $\leq$ is a preorder on them.

**Definition 2.5 (Mixed epistemic contexts)** *A* mixed epistemic context *$c$ is a structure of the form $\langle \mathcal{W}, \mathcal{R}, \leq \rangle$, where $\mathcal{W}$ is a set of ordinary propositional formulas, $\mathcal{R}$ is a set of contributory rules, possibly hedged, and $\leq$ is a preorder on $\mathcal{R}$.*

We can encode the Drug scenario in the mixed context $c_{16} = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ where $\mathcal{W} = \{Perceive(R),\ Drug\}$, $\mathcal{R}$ includes the rules

$$r_{26} = \frac{Perceive(R) : \neg Drug}{Believe(R)} \quad \text{and}$$

$$r_{23} = \frac{Perceive(R) \& Drug}{Believe(B)},$$

and where the priority relation is empty. At this point we must ask the perpetual question: How are we to get to the $\bigcirc$-formulas from mixed contexts like $c_{16}$? The first step is to formulate a notion that's equivalent to those of admissible and binding rules. And here I simply combine the ideas behind them:

**Definition 2.6 (Optimal rules)** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a mixed context. The rules from $\mathcal{R}$ that are* optimal *in it are those that belong to the set*

$$\begin{aligned} Optimal(c) \ = \ \{r \in \mathcal{R} : \ & \textit{(i) } r \in Triggered(c), \\ & \textit{(ii) there is no } \neg Z \in Hedge[r] \textit{ such that } \mathcal{W} \vdash Z, \textit{ and} \\ & \textit{(iii) there is no } r' \in Triggered(c) \textit{ such that} \\ & \quad (1)\ r \leq r' \textit{ and} \\ & \quad (2)\ contrary_c(r, r')\}. \end{aligned}$$

And, again, we can simply plug this notion into the definition of consequence that was formulated back in Section 1.2.

**Definition 2.7 (Consequence, mixed)** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a mixed epistemic context. Then the statement $\bigcirc X$ follows from c just in case $X \in Conclusion[Optimal(c)]$.*

It's not difficult to see that $\bigcirc Believe(B)$ does, while $\bigcirc Believe(R)$ doesn't follows from the mixed context $c_{16}$, as desired: The only rule that qualifies as optimal is $r_{23}$. The rule $r_{26}$ is triggered in $c_{16}$ and so satisfies condition (i). There's no other rule in $\mathcal{R}$ that would be contrary to $r$, weightier, and triggered; and so $r_{26}$ satisfies condition (iii). However, the hedge of $r_{26}$ is entailed by the hard information $\mathcal{W}$, which means that it fails to satisfy condition (ii).

How are we to think about reasons and defeat in this mixed model? The first thing to note here is that we can easily distinguish between rebutting and undermining defeat, and in a way that may look more natural than the one we used in the case of our model of the (absolute) hedged-rules view.

**Definition 2.8 (Rebutting and undermining defeat, mixed)** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a mixed context and $r$ a rule from $\mathcal{R}$ that's triggered in it. Then:*

*(1) $r$ is* rebutted *by a consideration $Z$ in $c$ if and only if $Z = Premise[r']$ for some rule $r'$ such that $r' \in Triggered(c)$, $contrary_c(r, r')$, and $r \leq r'$;*

*(2) $r$ is* undermined *by a consideration $Z$ in $c$ if and only if $\neg Z \in Hedge[r]$ and $\mathcal{W} \vdash Z$.*

Accordingly, we can say that a consideration $X$ qualifies as a reason for $Y$ in the

context $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ when the contributory hedged rule that $X$'s being a reason for $Y$ depends on is triggered in $c$ and not undermined by any other consideration. Now, given this way of understanding reasons in the model, we get the intuitive verdict for $c_{16}$: Since the rule $r_{26}$ is undermined by the proposition $Drug$, its premise $Perceive(R)$ does not qualify as a reason for believing that the object is red. So all is good here.

Given that we were able to show that the (simple) contributory-rules view is equivalent to a restricted version of the hedged-rules view, one may wonder about the relations between the newly formulated mixed view and the *full* version of the hedged-rules view. The following two observations show that the two views are, again, equivalent—both observations are verified in the Appendix:[18]

**Observation 4** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a hedged context subject to the following constraint: For any two rules $r$, $r' \in \mathcal{R}$ such that $contrary_c(r, r')$, either $\neg Premise[r'] \in Hedge[r]$ or $\neg Premise[r] \in Hedged[r']$. Then there's a mixed context $c' = \langle \mathcal{W}', \mathcal{R}', \leq \rangle$ such that*

*(1) $\mathcal{W}' = \mathcal{W}$;*

*(2) for every rule $r' \in R'$, there's a counterpart rule $r \in R$;*

*(3) $X$ is a reason for $Y$ in $c$ if and only if $X$ is a reason for $Y$ in $c'$;*

[18] Extending the definition of rule counterparts to take into account the mixed view is straightforward: Let $r$ be a hedged rule of the form $\frac{X \,:\, \mathcal{Z}}{\bigcirc Y}$ and $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ a mixed context. If there's a rule $r' \in \mathcal{R}$ such that $Premise[r'] = X$ and $Conclusion[r'] = Y$, we say that $r'$ is the *counterpart* of $r$ in the context $c$, written as $counterpart_c(r) = r'$. Going in the other direction, let $r$ be a rule and $c = \langle \mathcal{W}, \mathcal{R} \rangle$ a hedged context. If there's a rule $r' \in \mathcal{R}$ such that $Premise[r'] = X$ and $Conclusion[r'] = \bigcirc Y$, we say that $r'$ is the (hedged) *counterpart* of $r$ in the context $c$, written, yet again, as $counterpart_c(r) = r'$.

*(4) $X$ is rebutted as a reason for $Y$ by a consideration $Z$ in $c$ if and only if $X$ is rebutted as a reason for $Y$ by $Z$ in $c'$;*

*(5) $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.*

**Observation 5** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a mixed context subject to the following constraint: For any two rules $r$, $r' \in \mathcal{R}$ such that $contrary_c(r, r')$, either $r \leq r'$ or $r' \leq r$. Then there is a hedged context $c' = \langle \mathcal{W}', \mathcal{R}' \rangle$ such that*

*(1) $\mathcal{W}' = \mathcal{W}$;*

*(2) for every rule $r' \in R'$, there's a counterpart rule $r \in R$;*

*(3) $X$ is a reason for $Y$ in $c$ if and only if $X$ is a reason for $Y$ in $c'$;*

*(4) $X$ is rebutted as a reason for $Y$ by a consideration $Z$ in $c$ if and only if $X$ is rebutted as a reason for $Y$ by $Z$ in $c'$;*

*(5) $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.*

What's more, I strongly suspect it should be possible to establish another correspondence result, that is, to show that our model of the mixed view is equivalent to Horty's model with its special type of exclusionary rules.[19] But whether or not this is indeed the case will have to be left for future work.

Where does this leave us with regard to various views on simple rules and cases involving undermining defeat? Well, it looks like the mixed view and the full

[19] I have already noted that Horty's framework is more expressive than the one I'm using. Hence, my claim asks for a qualification: I suspect that the model of the mixed view is equivalent to a fragment of Horty's model.

version of the hedged-rules view can adequately deal with such cases, while the contributory-rules view and the restricted version of the hedged-rules view—that corresponds to it—cannot adequately deal with them. But does this also mean that the former views are overall superior to the latter? Here the matter is less clear. When we formulated the hedged-rules view back in Section 1.3.2, we also discussed a worry about it: In case it turns out that rule hedges are *open-ended*, or that there are infinitely many types of descriptive features that would make the instances of Testimony, Perception, and other rules schemas fail to apply, then the hedged-rules view may well reduce to the unattractive Über-rule view. Unfortunately, I don't know how to respond to this worry, and so it lingers as a potential threat to the full hedged-rules view and the mixed view.

The contributory-rules view and the restricted hedged-rules view, on the other hand, do have a ready response to the worry. According to the former, rules don't even have hedges, and so it's not even clear if the worry applies. And while one could suggest that the view might still reduce to the Über-rule position, if it turned out that we need infinitely many simple rule *schemas* to account for each and every epistemic situation, this though just doesn't seem to be plausible.[20] So, if there's a way to deal with the problem that cases of undermining defeat create—perhaps, one can argue that there simply isn't ay need for residual reasons in epistemology and change the formal notion accordingly—then the contributory-rules view would be on safer ground than either the full hedged-rules view or the mixed view. And so, in

[20]See Bradley (2019) and Holton (2002) who suggest that a view on which there are infinitely many rules reduces to something like an Über-rule view. See also McKeever & Ridge (2006) who argue that those authors who have tried compiling lists of basic moral principles have ended up with rather short lists.

fact, would be its hedged counterpart. How come? Well, supposed that the number of simple epistemic rule schemas is finite, we can be sure that their hedges will also be manageable. For, in light of the correspondence result from Section 2.1, we know that any rule that fails to apply in a situation doesn't apply because of a descriptive features that *also* triggers some other rule. But if there's a finite number of simple rule schemas, then the number of types of such features will be finite too. From here, the list of exceptions to any rule schema will be finite too, and not open-ended.

All in all, the contributory-rules view and the restricted hedged one appear to have no place for undermining defeat, but have a ready response to the particularist challenge about the prospect of the complexity of the view spiraling out of control. The mixed view and the full hedged view can account for undermining defeat, but are still in need of a convincing response to the challenge.

## 2.3 Summary

We started with the observation that situations of epistemic conflict give rise to a serious challenge for the idea that there are at least some simple epistemic rules. In response to this challenge, one can adopt a view on which such rules have built-in hedges, a view on which they are contributory (rather than strict), or a view that combines the first two. We expressed these three seemingly very different responses to the challenge in a simple framework and went on to establish some correspondence results between them. The (unhedged) contributory-rules view turns out to be equivalent to a restricted version of the (absolute) hedged-rules view, and the mixed

view, on which rules are both hedged and contributory, turns out to be equivalent to the full version of the (absolute) hedged-rules view. These correspondence results do not only establish that the views are extensionally equivalent, but also that they are connected at a deeper level. The simpler views have trouble adequately accounting for undermining defeat, but have a ready response to the particularist worry about the view's complexity. The more complex views can account for such defeat, but they will also have to provide an adequate response to the challenge.

# Part II DEFEASIBLE REQUIREMENTS

# Chapter 3: Misleading higher-order evidence and conflicting ideals

This second part of the dissertation has two goals. The first is to discuss a further application of the model that we used to capture the conciliatory-rules view: After a slight amendment, it can be used to develop a solution to an important puzzle about epistemic rationality. This solution has a number of attractive features when compared to the other solutions from the literatures, but it also comes with an unorthodox perspective on rationality requirements, one on which they are defeasible. The second goal, then, is to situate this perspective in the existent literature. More specifically, I show that we can naturally make sense of defeasible epistemic requirements as (regulative) epistemic ideals, and that the defeasible logic we use to solve the puzzle can be naturally seen as the formal backbone of Christensen's (2007, 2010, 2013) *conflicting-ideals view*. In effect, I'm proposing to understand this view as a move away from the default metaepistemological position according to which epistemic requirements are strict and governed by a strong, but never explicitly stated logic, toward the more unconventional view, according to which requirements are defeasible and governed by a comparatively weak logic.

The puzzle in question involves a clash between two widely-accepted requirements of rationality.[1] According to the first—the *evidential requirement*—epistemic rationality requires that you have the doxastic attitudes that are supported by your (total) evidence. According to the second—the *inter-level coherence requirement*—

[1]This puzzle has been noticed independently by several authors, including Lasonen-Aarnio (2020), Littlejohn (2018), and Worsnip (2018).

rationality requires that your doxastic attitudes are in line with your beliefs about whether or not these attitudes are supported by your evidence. We will state the two requirements more precisely in the next section, but for now simply note that they are distinct, intuitively plausible, and have been supported by strong arguments.[2] The trouble is that they come into conflict if we think—as I think we should—that it's possible for evidence to be radically misleading regarding what it itself supports. That is:

> **Misleading total evidence (MTE)**: It's possible that (1) your total evidence supports some doxastic attitude; and that (2) your total evidence supports believing that your total evidence doesn't support this doxastic attitude.[3]

The best, if still contested, illustrations of this sort of evidence from the literature are cases in which agents receive superb, yet misleading higher-order evidence. Here's one example:

> **Prof. Moriarty's Drug**: Sherlock Holmes is a master sleuth famously good at assessing evidence. After investigating a murder that took place at the manor on the hill he concludes (correctly) that the maid is the one who did it. Not long afterwards Holmes finds out, from a very reliable source, that his archnemesis Professor Moriarty had slipped a drug into his morning tea. Holmes knows the drug's effects too well: In all but 5%

[2]For a defense of the evidential requirement see (Conee & Feldman 2004) and (Lasonen-Aarnio 2020), among others. For extended defenses of the inter-level coherence requirement see (Feldman 2005), (Horowitz 2014), (Lasonen-Aarnio 2020), (Sliwa & Horowitz 2015), and (Worsnip 2018), among others.

[3]The label "Misleading total evidence" comes from (Worsnip 2019).

> of detectives, the ability to assess evidence gets distorted, and in a way that is not noticeable to them. What Holmes doesn't know is that he is among the lucky 5% that are immune to the drug.[4]

Advocates of (MTE) are pulled to the idea that Holmes' first-order evidence—roughly, the clues lying around the house—is sufficient evidence to believe that the maid did it, and that his higher-order evidence—roughly, the information speaking to his capacity for evaluating clues—is sufficient evidence to believe that his (total) evidence does not support believing that the maid did it. Thus, Moriarty's Drug would seem to be a case in which both (MTE1) and (MTE2) obtain.[5]

Once this much is granted, we quickly get into trouble: Holmes' evidence supports believing that the maid did it, as well as believing that his evidence doesn't support believing that she did. In light of this, the evidential requirement appears to demand that Holmes believes that the maid did it, as well as that his evidence doesn't support believing that she did. The coherence requirement, on the other hand, appears to demand that Holmes doesn't believe that the maid did it if he also believes that his evidence doesn't support believing that she did. A minute's reflection, then, reveals that were Holmes to comply with the first requirement, he

[4]This vignette amalgamates cases from (Christensen 2007a) and (Coates 2012).

[5]For the most forceful arguments in favor of (MTE) see (Lasonen-Aarnio 2020, Sec. 3) and (Worsnip 2018, 2019); see also (Coates 2012), (Littlejohn 2018), and (Weatherson ms). It's been suggested to me that (MTE) is more controversial than my setup of the puzzle suggests, as multiple authors have argued against it. But while (MTE) may be less intuitive of a starting point than the two requirements, I've yet to see an argument against it that wouldn't be motivated by a prior commitment to one of the requirements. As a result, I think of those arguing against (MTE) as responding to the puzzle by denying it—see footnote 19 for a list of references. Also, it's worth highlighting that the solution I'm going to present should be of interest independently of the particular puzzle—and thus the truth of (MTE)—as it naturally generalizes to *any* puzzle involving a conflict between epistemic requirements.

would thereby violated the second, and the other way around. So, supposing the two requirements are indeed genuine, Holmes simply can't do what epistemic rationality requires him to do, implying that Moriarty's Drug describes a rational dilemma.

What are we to say in response to this? That one of the requirements isn't genuine? That total evidence can never radically mislead about itself? That there are two independent domains of epistemic rationality, a domain that concerns having attitudes that track one's evidence and a domain concerned with having the right sort of fit between one's attitudes? Or, perhaps, simply bite the bullet and concede the existence of dilemmas? Well, although these responses have been explored in the literature, all of them incur intuitive costs and no single one has emerged as a clear winner.[6] And so we are still left with a genuine puzzle about rationality, a puzzle that has drawn and continues to draw much attention, in epistemology and outside it.

The main goal of this chapter is to present a solution to this puzzle which—unlike any of the others—lets us preserve the two requirements, the possibility of radically misleading (total) evidence, as well as the unity of epistemic rationality, all while denying that there are dilemmas. There's a sense in which this solution is implicit in Christensen's conflicting-ideals view. (According to this view, there are inherently unfortunate epistemic situations in which the agent is unable to act in accordance with all the "epistemic ideals" that apply to them, and not due to their cognitive limitations, but, rather, the way these ideals relate to each other.[7]) But the

[6]I should say upfront that I will not engage with these responses in what follows, focusing on developing my preferred solution instead. I will, however, provide a fuller classification of responses with pointers to the literature in Section 3.1.2.

[7]Various tragic epistemic scenarios of this sort are discussed in, e.g., (Christensen 2007a, 2013,

route that will get us to the solution is both new and nonstandard. What's more, taking it will help us understand the conflicting-ideals view itself much better.[8] The core idea is this: There's an implicit assumption in the literature about what we might call the *logic of interaction between requirements*, and the disconcerting result—that is, the commitment to dilemmas—follows only if this assumption is in place. The conflicting-ideals view, then, can be naturally thought of as rejecting this assumption and holding that this logic is weaker than standardly thought, and that rationality requirements are, in fact, defeasible.

The remainder of this chapter is structured as follows. Section will 3.1 restate the puzzle in a formal notation similar to the one we used in the previous chapters. This exercise will help us unearth the hidden assumption, as well as put us in a position to apply (an enhanced version of) the model we defined back in Section 1.4 to solve the puzzle—this will be the task of Section 3.2. As we'll see, there's good reason to think of this model as the formal backbone of the conflicting-ideals view.

## 3.1 The puzzle formalized

### 3.1.1 Preliminaries

Before we turn to the notation, three notes are in order. First, following Alex Worsnip (2018, 2019), we'll be thinking of *evidential support* as a two-place relation

2016), (Leonard 2020), (Pryor 2018), and (Schechter 2013).

[8]It's noteworthy that the literature focusing on the puzzle usually takes Christensen to hold a view that concedes the existence of rational dilemmas. See e.g., Lasonen-Aarnio (2020), Silva (2017), and Worsnip (2018) all of whom suggest that he embraces dilemmas. All three authors propose it as *one* interpretation of Christensen's views, but no other interpretation is ever discussed.

that obtains between bodies of evidence and doxastic attitudes. It's worth highlighting that this way of thinking comes with a certain sort of agent-neutrality: Once you fix the body of evidence, the doxastic attitudes it supports are also fixed. This doesn't mean that we're doing away with the agent—any body of evidence will be some agent's body of evidence—but it does mean that we can suppress the agent in the notation. Second, we'll try to remain as noncommittal as possible about the nature of evidence and evidential support relation. What I'm going to say should be compatible with various ways of thinking about evidence—as a set of propositions, facts, or mental states—as well as various ways of spelling out the support relation. And third, we'll restrict attention to all-or-nothing attitudes—this is not to take a stance in the debate about the relation between credence and full belief, but to keep things manageable.[9]

Our background language will, again, be the language of propositional logic with all the standard connectives. We supplement the language with the constant $E$ that will stand for an agent's total body of evidence. (Sometimes $E$ will stand for a particular body of evidence, and at other times for an arbitrary one—the context will always let disambiguate.) Further, we will make use of three operators, two of which we have seen before: The one-place $Believe(\cdot)$ will be used to capture all-or-nothing doxastic attitudes. Thus, $Believe(P)$ says that the agent believes $P$, and $Believe(\neg Q)$ says that the agent disbelieves $Q$.[10] The two-place operator

[9] As is known too well, issues surrounding misleading higher-order evidence do not disappear once one moves to the credence-based framework—see e.g., (Christensen 2010a). So sticking with the more parsimonious all-or-nothing attitude-based framework seems like a methodologically sound move.

[10] We can also easily express suspension of judgment, either as $\neg Believe(P) \& \neg Believe(\neg P)$, or by introducing a new predicate $Suspend(\cdot)$. But this particular epistemic attitude won't have

$\Mapsto$, in turn, will be used to formulate claims about the evidential support relation. Thus, $E \Mapsto \mathit{Believe}(P)$ says that the body of evidence $E$ supports believing $P$, and $E \Mapsto \mathit{Believe}(\neg Q)$ says that $E$ supports disbelieving $Q$. Note that $\Mapsto$ is meant to stand for all things considered support. When an attitude $\mathit{Believe}(X)$ is not supported by the evidence, we will write $E \not\Mapsto B(X)$. We will also form more complex expressions, including formulas that capture second-order beliefs—an agent's beliefs about whether a certain attitude is supported by their evidence—as well as formulas that assert the existence of support between $E$ and such beliefs. For example, $\mathit{Believe}(E \Mapsto B(P))$ says that one believes that one's evidence supports believing $P$; and $E \not\Mapsto \mathit{Believe}(E \Mapsto \mathit{Believe}(P))$ says that the evidence doesn't support having this second-order belief. Finally, we'll make use of the customary deontic operator $\bigcirc(\cdot)$. The requirements of epistemic rationality set up a certain epistemic standard, and so a statement of the form $\bigcirc X$ should be read as saying that, according to the standards of epistemic rationality, it ought to be the case that $X$. In what follows, we will often refer to this standard as the *epistemic ought*.

With this notation in hand, the puzzle can be brought into plain sight. It's comprised of three claims. The first captures the evidential requirement:

**Evidential requirement (ER)**: $[E \Mapsto \mathit{Believe}(X)] \supset \bigcirc \mathit{Believe}(X)$

In English: If your evidence supports believing $X$, then you ought to believe $X$. This thesis is very intuitive, with many epistemologists considering it a platitude.[11] Also, it—or something close to it—lies at the heart of the popular philosophical

much of a role to play in the discussion.

[11] Cf. (Silva 2017) and (Worsnip 2018).

view called *evidentialism*.[12] The second claim captures the coherence requirement:

**Inter-level coherence requirement (ILC)**:

(1) $\bigcirc[Believe(E \Rightarrow B(X)) \supset Believe(X)]$

(2) $\bigcirc[Believe(E \not\Rightarrow Believe(X)) \supset \neg Believe(X)]$

In English: (1) Rationality requires that you believe a proposition if you believe that your evidence supports that belief. (2) Rationality requires that you do not believe a proposition if you believe that your evidence doesn't support that belief. Much of the appeal of (ILC) derives from the intuition that there's something seriously wrong with agents who believe, act, or assert in accordance with mismatched attitudes of the sort (ILC) prohibits. Note the difference in the formal statement of (ER) and (ILC). In the former the ought occurs in the consequent, while in the latter it ranges over the entire conditional. This reflects the distinction between narrow- and wide-scope rationality requirements that goes back at least to (Broome 1999).[13] Assigning (ER) and (ILC), respectively, narrow and wide scope is standard in the literature.[14]

[12]In fact, we can define evidentialism via (ER): An agent is epistemically rational if and only if their attitudes satisfy (ER). For a concise introduction to evidentialism see (Conee & Feldman 2008). For book-length defenses see (Conee & Feldman 2004) and (McCain 2014).

[13]Narrow-scope requirements demand that a particular attitude is adopted once the conditional's antecedent is satisfied. Wide-scope requirements are supposed to give the agent more freedom. Consider Alice who finds herself believing that her evidence supports a belief in $p$. There are two ways for her to comply with (ILC1)—and we're supposing that complying with it is all that matters. Alice can either adopt a belief toward $p$, or she can drop her initial belief. For more on narrow vs. wide scope see, e.g., (Broome 2007, 2013). For a recent criticism of the distinction see (Fogal f).

[14]See e.g., (Lasonen-Aarnio 2020), (Silva 2017), (Worsnip 2018). Those familiar with the debate about the normativity of rationality might have the following thought here: On some views in the debate, such structural requirements as (ILC) are either nonnormative—see e.g., (Kolodny 2005)—or simply nonexistent—see (Kiesewetter 2017) and (Lord 2018). So the proponents of such views wouldn't accept (ILC) and, thereby, also avoid the puzzle. But while they would avoid the particular puzzle we focus on here, they have to deal with a closely related one. Here's why. Denying (ILC) standardly goes in hand with accepting disjunctions of (non-

The third and final claim is the possibility of total evidence that's radically misleading regarding itself. We have noted already that many epistemologists would characterize Holmes' evidence in Moriarty's Drug as evidence of just this sort. That is, they would say that his total evidence supports both believing that the maid did it, and believing that his evidence doesn't support believing that she did. Letting $M$ stand for the proposition that the maid committed the murder, we acquire:

**Misleading total evidence (MTE)**:

(1) $E \Rightarrow \mathit{Believe}(M)$

(2) $E \Rightarrow \mathit{Believe}(E \not\Rightarrow \mathit{Believe}(M))$

Note that Moriarty's Drug is only an example, and that we have a puzzle in case it's metaphysically possible that there's *some* body of evidence $E$ and *some* proposition $X$ with both $E \Rightarrow \mathit{Believe}(X)$ and $E \Rightarrow \mathit{Believe}(E \not\Rightarrow \mathit{Believe}(X))$.[15]

### 3.1.2 Deriving the disconcerting result

Now all the pieces are in place, and we can construct a proof showing how exactly (ER), (ILC), and (MTE) lead to the conclusion that Moriarty's Drug describes a dilemma. We start with facts describing evidential support:

(1) $E \Rightarrow \mathit{Believe}(M)$ (MTE1)

structural) requirements $\bigcirc\neg\mathit{Believe}(E \Rightarrow \mathit{Believe}(X)) \vee \bigcirc\mathit{Believe}(X)$ and $\bigcirc\neg\mathit{Believe}(E \not\Rightarrow \mathit{Believe}(X)) \vee \bigcirc\neg\mathit{Believe}(X)$, and, when these are combined with the other components of the puzzle, very similar problems arise. For a discussion of the kindred puzzle see (Kiesewetter 2017, Ch. 9.6.2).

[15] Cf. (Lasonen-Aarnio 2020) and (Worsnip 2018).

(2) $E \Mapsto Believe(E \not\Mapsto Believe(M))$ (MTE2)

These determine the relevant instances of (ER):

(3) $[E \Mapsto Believe(M)] \supset \bigcirc Believe(M)$ (ER)

(4) $[E \Mapsto Believe(E \not\Mapsto Believe(M))] \supset \bigcirc Believe(E \not\Mapsto Believe(M))$ (ER)

Two applications of modus ponens result in (5) and (6)—which say, respectively, that Holmes ought to believe that the maid did it, and that Holmes ought to believe that the evidence doesn't support believing that the maid did it:

(5) $\bigcirc Believe(M)$ from (1) and (3)

(6) $\bigcirc Believe(E \not\Mapsto Believe(M))$ from (2) and (4)

Next comes the following instance of (ILC):

(7) $\bigcirc[Believe(E \not\Mapsto Believe(M)) \supset \neg Believe(M)]$ (ILC2)

This formula says that Holmes ought not to believe that the maid did it in case he believes that the evidence doesn't support believing that the maid did it. The subsequent step (8) is the crucial for our purposes. To get further, we need to appeal to a logical feature of the deontic operator. More specifically, we need to assume that it satisfies the principle $\bigcirc(X \supset Y) \supset (\bigcirc X \supset \bigcirc Y)$.[16] Instantiating $X$ with $Believe(E \not\Mapsto Believe(M))$ and $Y$ with $\neg Believe(M)$, we get (8):

[16]In the logic literature this is known as the *axiom K*, and it's a very natural principle for a deontic operator to satisfy. But see (Broome 2013, Ch. 7) for an argument to the conclusion that a logic of requirements (and oughts) shouldn't be closed under $K$.

(8) $\bigcirc[Believe(E \nRightarrow Believe(M)) \supset \neg Believe(M)] \supset$

$(\bigcirc Believe(E \nRightarrow Believe(M)) \supset \bigcirc\neg Believe(M))$ (Deontic Principle)

Then we have two more applications of modus ponens:

(9) $\bigcirc Believe(E \nRightarrow Believe(M)) \supset \bigcirc\neg Believe(M)$ from (7) and (8)

(10) $\bigcirc\neg Believe(M)$ from (6) and (9)

And, as a final touch, we combine the formulas acquired in steps (5) and (10):

(11) $\bigcirc Believe(M) \& \bigcirc\neg Believe(M)$ conjunction, from (5) and (10)

This last formula is disconcerting because it entails that Holmes finds himself in a normative dilemma, or a situation in which it ought to be the case that $X$ and it ought to be the case that $Y$, while it is impossible for both $X$ and $Y$ to obtain.[17] If Holmes complies with either ought on line (11), he thereby violates the other. So there's no way he can have the attitudes he ought to have. And given that the ought statements express demands of epistemic rationality, the derivation suggests that there are rational dilemmas.

Most epistemologists have tried to resolve the puzzle in ways that keep clear of dilemmas.[18] In particular, they have suggested rejecting (MTE),[19] rejecting (ER),[20]

[17]This is the standard definition of normative dilemmas—see e.g., (Goble 2009, p. 450) and (Horty 2003, p. 557).

[18]The list of authors most likely to be sympathetic to embracing the derivation's conclusion would include Hughes (2017) and Priest (2002).

[19]See (Feldman 2005), (Horowitz 2014), (Kiesewetter 2017, Ch. 9), (Skipper 2019), (Tal f), (Titelbaum 2015), (White 2007), among many others.

[20]See (Littlejohn 2018).

rejecting (ILC),[21] or treating the conflict as a clash between two irreducible types of epistemic rationality.[22] The derivation presented here provides a way to classify these responses: The first move denies one of the premises, the second denies either step (3) or (4), the third denies step (7), and the fourth rejects step (10).[23]

The real significance of the derivation, however, consists in it explicitly showing that the normative component picked out by $\bigcirc$ makes its own contribution to the puzzle. In order to get to the disconcerting result, we had to rely on a logical feature of $\bigcirc$. On the intuitive level, this can be understood thus: There's a nontrivial logic governing the interaction between epistemic requirements, and the puzzle causes real trouble only if it is further assumed—as most of the literatures has done hitherto—that this logic is relatively strong.

But this assumption can and should be questioned, as one attractive and underexplored route of response to the puzzle is to weaken the logic. And I'm suggesting that we understand the conflicting-ideals view as taking just this route. Let me emphasize that I am *not* suggesting that it simply denies that the epistemic ought satisfies the deontic principle we used in the proof. Instead, my suggestion is that we understand the view as not only changing the logic governing the interaction between epistemic requirements, but also distinguishing it from the logic of the

[21] Authors who are often taken to reject (ILC) include Coates (2012), Lasonen-Aarnio (2020), and Weatherson (ms). But Lasonen-Aarnio appears to be the only one to explicitly advocate the rejection of (ILC) in response to the puzzle.

[22] See (Fogal f), (Silva 2017, especially, en. 11), and (Worsnip 2018).

[23] How exactly does it reject step (10)? Well, two types of rationality means two distinct normative domains, and, hence, also two distinct deontic operators. Worsnip, for instance, distinguishes between the domains of evidence and coherence, and so implicitly also between the the oughts of evidence and coherence. The ought occurring on line (6) is an ought of evidence, while the one occurring on line (9) is an ought of coherence. Since the oughts are distinct, we can't apply modus ponens.

epistemic ought. The result is a view of epistemic normativity, according to which epistemic oughts are determined through the interaction of *defeasible* epistemic requirements, or *ideals*. This response to the puzzle is hardly ad hoc. For, first, it naturally generalizes to other puzzles involving conflicts between requirements. And second, it parallels a familiar and well-respected move in the ethical literature on moral conflicts.

What we are going to do next is define a concrete defeasible deontic logic as a substitute for the implicitly assumed logic of $\bigcirc$ and, then, see how it can help us avoid the disconcerting result. As flagged above, I see this logic as the formal backbone of the conflicting-ideals view.

## 3.2 A (formal) solution

### 3.2.1 A logic for conflicting ideals

The logic we'll use isn't really new. Its core is a well-known and studied defeasible consequence relation defined in terms of classical consequence and maximally consistent subsets—I'll explain the latter notion in due time. Nicholas Rescher and Ruth Manor (1970) appear to have been the first to define and study consequence relations of this sort, and Horty (1994, 2003) was the first to apply them in deontic setting, in the context of the debate over the existence of moral dilemmas. They also have close connections to Bas van Fraassen's (1973) deontic logic, Raymond Reiter's (1980) default logic, and other logics from the defeasible logic paradigm.[24] The con-

[24]See Horty (1994) and Makinson (2005).

sequence relation we'll define here is a close cousin of what's called the *disjunctive account* in (Horty 2003). For expository purposes, we will discuss its application not only in the epistemic, but also in the moral domain.

The first thing we need to do is distinguish between a weaker and a stronger sense of *ought*. This move is standard in the literature on moral dilemmas: Thus, Roderick Chisholm (1964) writes about *prima facie* and *absolute duties*, van Fraassen (1973) about *imperatives* and *oughts*, John Searle (1980) about *obligations* and *oughts*, Philippa Foot (1983) about *type 1* and *type 2 oughts*, Christopher Gowans (1987) about what we *ought* and what we *must* do, and Horty (2003) simply about *weak* and *strong oughts*. The terms used are different, but the underlying idea is always the same: Take some situation that requires a nontrivial response on the part of the agent. The first term of the pair would be used to pick out various moral considerations the agent has and the responses they support; and the second term would be used to describe what the agent's response ought to be once all the relevant considerations are taken into account and weighed against each other. For illustration consider the well-known example from Sartre (1946) in which a French youth at the time of Second World War is torn between the patriotic duty to fight for his country and the filial duty to care for his distressed and aging mother. All the parties to the debate about moral conflicts would agree—at least, once they left the terminological differences behind—that there's a weaker sense of *ought*, according to which the youth ought to fight for his country and also ought to stay with his mother. What they wouldn't agree on is whether the youth ought to do both of these things in the stronger sense of *ought* too—that is, whether he faces a genuine

moral dilemma.

While drawing an analogous distinction between a weaker and a stronger sense of the *epistemic ought* is less common, nothing stands in the way of doing it. What's more, if I am right, the conflict-ideals diagnosis of the various inherently unfortunate or tragic epistemic scenarios discussed in the recent literature entails such a distinction, as well as the claim that ideals (or requirements) that come into conflicts in such scenarios express the weaker epistemic oughts.[25]

Since our main interest is in epistemology, we will employ terminology that is a better fit for the epistemic than the moral domain: We will refer to the weaker oughts—whether they be epistemic or moral—as (defeasible) *requirements* or *ideals* and to the stronger ones as *all things considered oughts*, or *oughts* without qualification. In the ethical literature, it is common to take the stronger oughts to be generated from the weaker ones. And that's just how our logic will work: It will generate all things considered oughts from the interaction between requirements.

Reserving the $\bigcirc$ for the oughts, we will express requirements in the same way we expressed rules in the previous chapters. Where $X$ and $Y$ are arbitrary propositions, $\frac{X}{Y}$ says that there's a demand that $Y$ obtains under circumstances $X$, or, alternatively, that, ideally, $Y$ should obtain under circumstances $X$. Importantly, this demand can get overridden. We will denote requirements using the letter $r$ (with subscripts). It will, again, be useful to have functions for picking out the two parts of a requirement: Where $r$ stands for the requirement $\frac{X}{Y}$

[25]Let me add a note of caution: Since the moral ought in its weaker sense is roughly synonymous with the widely used notion of a moral reason, you might naturally expect that the epistemic ought in its weaker sense must be roughly synonymous with the notion of an epistemic reason too. However, in the present context the latter two are definitely not synonymous.

, the output of $Premise[r]$ is the proposition $X$ and that of $Conclusions[r]$ is the proposition $Y$. Just like it was in the case of rules, $Premise[r]$ specifies the *triggering conditions* of $r$, or the circumstances under which it comes into force, while $Conclusion[r]$ specifies its *satisfaction conditions*, or the circumstances under which the requirement's demand gets fulfilled. We will, again, lift the second function from individual requirement to sets of them: Where $\mathcal{R}$ is a set of requirements, $Consequent[\mathcal{R}]$ is the collection of the conclusions of all the requirements from $\mathcal{R}$, or $Consequent[\mathcal{R}] = \{Consequent[r] : r \in \mathcal{R}\}$.

Oughts will be generated from *contexts*, or triples of the form $\langle \mathcal{W}, \mathcal{R}, \leq \rangle$, consisting of a set of propositional formulas $\mathcal{W}$, a set of requirements $\mathcal{R}$, and a priority relation $\leq$ over $\mathcal{R}$. As a reminder, the relation $\leq$ satisfies reflexivity and transitivity. Thus, for all requirements $r$, $r'$, $r''$, we have $r \leq r$, as well as, if $r \leq r'$ and $r' \leq r''$, then $r \leq r''$. The expression $r < r'$ means that $r \leq r'$ and not $r \leq r'$, and $r \leq \mathcal{Q}$ that every requirement in the set of requirements $\mathcal{Q}$ has at least as much weight as $r$. I should also note that the shape of $\mathcal{W}$ will differ, depending on the scenario we are modeling: In case the scenario pertains to the moral domain, this set will contain formulas expressing the descriptive facts of the situation; in case the scenario pertains to the epistemic domain, it will contain formulas expressing facts about evidential support.

**Definition 3.1 (Contexts)** *A* context *$c$ is a structure of the form $\langle \mathcal{W}, \mathcal{R}, \leq \rangle$, where $\mathcal{W}$ is a set of propositional formulas, $\mathcal{R}$ is a set of requirements, and $\leq$ is a preorder on $\mathcal{R}$.*

Notice that what we call *contexts* here are the same mathematical objects as the *weighted contexts* we introduced back in Section 1.4.1. (We can drop the modifier since all the contexts we will encounter in this chapter are weighted.) In light of this, it shouldn't come as a surprise that we will be reusing some of our definitions. Before we turn to them, however, an example of a context is in order, a context representing a scenario we have encountered before:

> **Drowning Child**: You have promised your friend Taylor to have a dinner with her. Your route to the restaurant takes you past a pond, and, as you are walking past it, you notice that a child has fallen in. The child is crying in distress, and all your evidence suggests that it is going to drown, unless you do something about it. However, if you rescue the child, you will get your clothes wet and muddy, and won't make it to the dinner with Taylor.

Let $Promise_1$ and $Dine_1$ stand for the propositions, respectively, that you have a made promise to dine with your friend, and that you dine with her, and let $Drowning$ and $Save$ stand for the propositions that a child is drowning, and that you save the child. The requirement $r_1 = \frac{Promise_1}{Dine_1}$ would then express the demand that you dine with Taylor in case you've promised to dine with her, and $r_2 = \frac{Drowning}{Save}$ the demand that you save the child, given the need. We can encode this scenario into the context $c_1 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ where $\mathcal{W}$ contains $Promise_1$, $Drowning$, as well as the formula $\neg(Dine_1 \& Save)$, which captures the constraint that you can't both keep your promise and save the child; where $\mathcal{R}$ contains the

requirements $r_1$ and $r_2$; and where the second requirement takes priority over the first, or $r_1 < r_2$.

The next two definitions will specify how we are to select those requirements that are in force in a context. The first simply restates the definition of triggered rules, and the second a generalizes the definition of binding rules from Section 1.4.1.

**Definition 3.2 (Triggered requirements)** *Where $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ is a(n epistemic) context, the requirements from $\mathcal{R}$ that are* triggered in $c$ *are those that belong to the set $Triggered(c) = \{r \in \mathcal{R} : \mathcal{W} \vdash Premise[r]\}$.*

**Definition 3.3 (Binding requirements)** *Where $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ is a context, the requirements that are* binding in $c$ *are those that belong to the set*

$$
\begin{aligned}
Binding(c) \quad = \{r \in \mathcal{R} : \quad & (i)\ r \in Triggered(c)\ \textit{and} \\
& (ii)\ \textit{there's no } \mathcal{Q} \subseteq Triggered(c)\ \textit{such that} \\
& \quad (1)\ r \leq \mathcal{Q}, \\
& \quad (2)\ \mathcal{W} \cup Conclusion[\mathcal{Q}] \vdash \neg Conclusion[r],\ \textit{and} \\
& \quad (3)\ Conclusion[\mathcal{Q}]\ \textit{is consistent with } \mathcal{W}\}.
\end{aligned}
$$

In English: To enter the set $Binding(c)$ a requirement $r$ must (i) be triggered, and (ii) there can't be a set of triggered requirements $\mathcal{Q}$ that are uniformly at least as good as $r$, jointly consistent (with $\mathcal{W}$), and such that they entail the negation of $r$'s conclusion. Although it might not be obvious at first sight, this definition really is a conservative generalization of that of binding rules. When discussing rules, we made the simplifying assumption that a(n epistemic conflict) is a conflict between

two rules. In the present context, this assumption can no longer be upheld, since the clash of requirements in the Holmes scenario is, in fact, a clash between three instances of requirements. For this reason, clause (ii) in the definition refers to a set of requirements, instead of a single requirement that can preclude $r$ from qualifying as binding. Clause (ii:1) generalizes the expression $r \leq r'$; and clauses (ii:2) and (ii:3) generalize the notion of a contrary rule to the notion of a contrary set of rules. The need for (ii:3) will become manifest as soon as we apply the definition to a particular case, which is what we turn to next.

It's not difficult to see that both requirements $r_1$ and $r_2$ are triggered in the context $c_1$, while only the latter qualifies as binding. Notice how $r_1$ doesn't qualify because of the singleton set $\mathcal{Q}_1 = \{r_1\}$. First, it is a subset of $Triggered(c_1)$. Second, every requirement in it has at least as much weight as $r_1$, that is, $r_1 \leq \mathcal{Q}_1$ holds true. And third, $Conclusion[\mathcal{Q}_1] = \{Save\}$ is consistent with $\mathcal{W} = \{Promise_1, Drowning, \neg(Dine_1 \& Save)\}$, and together with $\mathcal{W}$ it entails $\neg Dine_1$, the negation of $Conclusion[r_1]$. Of course, the fact that $r_2$ does, while $r_1$ does not qualify as binding is the intuitive result. Now suppose that clause (ii:3) was not included in the above definition. In that case, $r_2$ wouldn't qualify as binding either: For there is a set of requirements that are triggered in $c_1$ that satisfies clauses (ii:1) and (ii:2), namely, $\mathcal{Q}_2 = \{r_1, r_2\}$.

The next question we must ask is how to get to all things considered oughts. Intuitively, they should be determined by the satisfaction conditions of the requirements that are in force, or binding in the context—in the particular case at hand by the satisfaction conditions of $r_2$. Once we recall that $Conclusion[\cdot]$ tells us just what

those satisfaction conditions are, the following definition should look very natural:

**Definition 3.4 (Consequence, first pass)** *Where $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ is a context, the all things considered ought statement*

$\bigcirc X$ *follows from c if and only if* $\mathcal{W} \cup Conclusion[Binding(c)] \vdash X$.

This definition generalizes the corresponding definition from the section on weighted rules, and it gives us the intuitively correct result for $c_1$. Since $Conclusion[\{r_2\}]$ equals $\{Save\}$ and $\mathcal{W} \cup \{Save\}$ entails $Save$ and $\neg Dine_1$, the formulas $\bigcirc Save$ and $\bigcirc \neg Dine_1$ follow from the context. It turns out, however, that we need to amend this definition if the logic is to deal with scenarios involving *conflicts* between requirements that are either incommensurable, or of the same weight. This becomes manifest once we try to apply the definition to another toy example from the practical domain:

> **Twins**: Your friend Taylor has a twin Tyler you are also good friends with. You have inadvertently promised to have a private dinner with each of the twins at the same time. Both Taylor and Tyler are equally important to you, and both would be equally disappointed by your cancellation.[26]

Let $Promise_1$ and $Dine_1$ be as before, and let $Promise_2$ and $Dine_1$ stand for the propositions, respectively, that you have promised to dine with the second twin Tyler and that you dine with him. The case itself can be encoded in the context $c_2 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ where $\mathcal{W} = \{Promise_1, Promise_2, \neg(Dine_1 \& Dine_2)\}$; where $\mathcal{R}$ contains

[26] The example is adapted from (Horty 2012, p. 30).

the requirements $r_1 = \frac{Promise_1}{Dine_1}$ and $r_3 = \frac{Promise_2}{Dine_2}$, with $r_1$ still saying that you are under a standing demand to dine with Taylor, if you've promised to dine with her, and $r_3$ saying that you are under a similar demand vis-à-vis Tyler, if you've made a promise to him; and where $\leq$ is empty. The formula $\neg(Dine_1 \& Dine_2)$, again, expresses the background constraint that it's impossible for you to dine with both twins. The two requirements $r_1$ and $r_3$ both get triggered in $c_2$, and both qualify as binding. And now there's a problem: Combining their satisfaction conditions with $\mathcal{W}$ in the way the definition requires results in an inconsistent set. Such sets entail all formulas, and so our definition has the counterintuitive consequence that the context $c_2$ entails a formula of the form $\bigcirc X$ for any $X$ whatsoever.[27]

What's at the root of the problem? Well, it's the fact that the definition requires that we use *all* the statements in $Conclusion[Binding(c)]$, a set that, in general, can be inconsistent (or inconsistent with $\mathcal{W}$). So we need some sort of restriction. And our strategy will be to fall back from the entire set $Conclusion[Binding(c)]$ to its largest consistent parts. The formal concept we'll be relying on is that of a maximally consistent, or maxiconsistent, subset:[28]

**Definition 3.5 (Maximally consistent subsets)** *Where $\mathcal{G}$ and $\mathcal{H}$ are two sets of propositional formulas, a subset $\mathcal{F}$ of $\mathcal{H}$, $\mathcal{F} \subseteq \mathcal{H}$, is said to be maxiconsistent with*

[27] One might be tempted to think that we have gotten to this absurd result because we didn't express the scenario correctly: Instead of not assigning any (relative) weights to the requirements $r_1$ and $r_3$, we should have assigned them the same weights. But, for better or worse, this alternative way of formalizing the scenario doesn't really solve the problem. If $r_1$ and $r_3$ have the same weight, then neither of them qualifies as binding, leading to the conclusion that there's no obligation for you to keep either of the promises, and this seems counterintuitive. Another claim one might make—in order to avoid the problem—is that requirements are never incommensurable and can also never have the same weight. But while this claim would let us avoid the problem, it is also in need of a substantial argument.

[28] More precisely, we're relying on a generalization of the concept from (Makinson 2005, p. 30ff).

*$\mathcal{G}$ if and only if (i) $\mathcal{F}$ is consistent with $\mathcal{G}$, and (ii) there is no consistent set $\mathcal{F}'$ such that $\mathcal{F} \subset \mathcal{F}'$ and $\mathcal{F}' \subseteq \mathcal{H}$.*

The concept is of most use when $\mathcal{G}$ and $\mathcal{H}$ are individually consistent, while $\mathcal{G} \cup \mathcal{H}$ isn't. Intuitively, a subset of $\mathcal{H}$ that's maxiconsistent with $\mathcal{G}$ is as big a subset of $\mathcal{H}$ as you can add to $\mathcal{G}$ without running afoul of inconsistency: Supplementing it with even one additional formula from $\mathcal{H}$ would render the result inconsistent—this is what clause (ii) ensures.

With this concept in hand, we can define which oughts follows from a given context $c$ by focusing *not* on the entire set $Conclusion[Binding(c)]$, but on those of its subsets that are maxiconsistent with $\mathcal{W}$. The plural is not accidental. An inconsistent set can have multiple maxiconsistent subsets, and our policy is to require that a statement $X$ follows from *all* such subsets if $\bigcirc X$ is to qualify as an all things considered ought. That is:

**Definition 3.6 (Consequence, final)** *Where $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ is a context, the all things considered ought statement $\bigcirc X$ follows from $c$ if and only if, for* every *subset $\mathcal{F}$ of $Conclusion[Binding(c)]$ that is maxiconsistent with $\mathcal{W}$, we have $\mathcal{W} \cup \mathcal{F} \vdash X$.*

Let's revisit our examples. First, notice that nothing has changed with regard to the context $c_1$, capturing the Drowning Child scenario. There's only one subset of $Conclusion[Binding(c_1)] = \{Save\}$ that's maximally consistent with $\mathcal{W}$, namely, $\{Save\}$ itself, and so $\bigcirc Save$ and $\bigcirc \neg Dine_1$ still follow from $c_1$. This illustrates the conservative character of the amendment. But what about $c_2$? As before, both requirements $r_1 = \dfrac{Promise_1}{Dine_1}$ and $r_3 = \dfrac{Promise_2}{Dine_2}$ qualify as binding, and we

get $Conclusion[Binding(c_2)] = \{Dine_1, Dine_2\}$. This set has two subsets that are maxiconsistent with $\mathcal{W}$, namely, $\{Dine_1\}$ and $\{Dine_2\}$. What is it, then, that you ought to do in the Twins case? Since neither $Dine_1$, nor $Dine_2$ follows from both $\{Dine_1, \neg(Dine_1 \& Dine_2)\}$ and $\{Dine_2, \neg(Dine_1 \& Dine_2)\}$, the context doesn't entail either $\bigcirc Dine_1$ or $\bigcirc Dine_2$. So it's not the case that you ought to dine with the first twin, nor is it the case that you ought to dine with the second. However, the disjunction $Dine_1 \vee Dine_2$ does follow from both of these sets, and so the context entails $\bigcirc(Dine_1 \vee Dine_2)$. This means that you can't just walk away; you have to keep one of your promises. There's also a sensible rationale behind this recommendation. If you comply with the ought—in either of the two ways—you keep as many promises as is humanly possible in your situation. You also break as few promises as you possibly can.

### 3.2.2 Back to the puzzle

With the logic at our disposal, we can return to the puzzle and Prof. Moriarty's Drug. We will discuss a couple of ways to capture it formally, starting with one that doesn't assign any weights to requirements. We will encode the scenario into the context $c_3 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$. Its first element $\mathcal{W}$ will contain the formulas expressing the relevant facts about evidential support, $E \Rightarrow Believe(M)$ and $E \Rightarrow Believe(E \not\Rightarrow Believe(M))$. Notice that this is the last component of the puzzle (MTE). The context's second element, in turn, will contain the relevant instances of the requirements (ER) and (ILC). In Section 3.1.2, we expressed them as

the strict requirements:

(3) $[E \Rightarrow Believe(M)] \supset \bigcirc Believe(M)$,

(4) $[E \Rightarrow Believe(E \not\Rightarrow B(M))] \supset \bigcirc Believe(E \not\Rightarrow Believe(M))$, and

(7) $\bigcirc[Believe(E \not\Rightarrow Believe(M)) \supset \neg Believe(m)]$.

Adapting (3) and (4) to the new setting, we acquire:

$$r_4 = \frac{E \Rightarrow Believe(M)}{Believe(M)} \text{ and}$$

$$r_5 = \frac{E \Rightarrow Believe(E \not\Rightarrow Believe(M))}{Believe(E \not\Rightarrow Believe(M)} .$$

The requirement $r_4$ says that, ideally, Holmes should believe that the maid did it in the circumstances where his evidence supports believing that she did. Similarly, the requirement $r_5$ says that, ideally, Holmes should believe that his evidence doesn't support believing that the maid did it in circumstances where his evidence supports this belief.

The requirement (ILC) is a little more tricky. We capture it as follows:

$$r_6 = \frac{\top}{Believe(E \not\Rightarrow Believe(M)) \supset \neg Believe(M)} .$$

The symbol $\top$ here stand for an arbitrary tautology; and given that $\top$ follows from any set of formulas, a requirement that has it in the antecedent is guaranteed to get triggered. So, on our rendering of (ILC), there's always a standing defeasible demand that one's second-order beliefs and first-order attitudes cohere with each other. It's worth highlighting that the rule $r_6$ is equivalent to $\frac{\top}{\neg(Believe(E \not\Rightarrow Believe(M)))\&Believe(M))}$

. This is also a very natural way to represent wide-scope requirements in our framework.[29]

All in all, then, we have the context $c_3 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ where $\mathcal{W}$ contains the expressions $E \Rightarrow Believe(M)$ and $E \Rightarrow Believe(E \not\Rightarrow Beliece(M))$, $\mathcal{R}$ contains the requirements $r_4$, $r_5$, and $r_6$, and $\leq$ is empty. But how does this help with solving the puzzle? Well, recall that the problem was that we were able to conclude that Holmes both ought to believe that the maid did it and ought to avoid having this belief, $\bigcirc Believe(M)$ and $\bigcirc \neg Believe(M)$. Now let's see what all things considered oughts follow from $c_3$. As a first step, notice that $r_4$, $r_5$, and $r_6$ all qualify as binding, and that $Conclusion[Binding(c_3)]$ has three subsets that are maximally consistent with $\mathcal{W}$, namely,

$\{Believe(M),\ Believe(E \not\Rightarrow Believe(M))\}$,

$\{Believe(M),\ Believe(E \not\Rightarrow Believe(M)) \supset \neg Believe(M)\}$, and

$\{Believe(E \not\Rightarrow Believe(M)),\ Believe(E \not\Rightarrow Believe(M)) \supset \neg Believe(M)\}$.

A minute's reflection reveals that the third set doesn't entail $Believe(M)$ and that the first doesn't entail $\neg Believe(M)$. This means that neither of the problematic oughts, $\bigcirc Believe(M)$ and $\bigcirc \neg Believe(M)$, follow from $c_3$. So it is not the case that Holmes ought to believe that the maid did, nor is it the case that he ought *not* to believe that the maid did it. However, as in the twins example, there's a disjunction that follows from all three sets, implying that the context $c_3$ entails the following (disjunctive) ought:

[29] Cf. (Schroeder 2018).

$$\bigcirc([Believe(M)\&Believe(E \not\Rightarrow Believe(M))] \quad \vee$$
$$[Believe(M)\&\neg Believe(E \not\Rightarrow Believe(M))] \quad \vee$$
$$[\neg Believe(M)\&Believe(E \not\Rightarrow Believe(M))]).$$

So Holmes can't adjust beliefs as he pleases; he has to do it in one of the specified ways. Unpacking the formula, he ought to either (i) believe both that the maid did it *and* that the evidence doesn't support believing that she did; or (ii) believe only that the maid did it; or (iii) believe only that the evidence doesn't support believing that the maid did it. Put differently, there are two relevant beliefs—that the maid did it and that the evidence doesn't support believing that she did—and Holmes is as he ought to be as long as he holds at least one of them.

Notice that this disjunctive recommendation can be supported by a rationale that parallels the one we appealed to in the twins case: By adjusting beliefs in *any* of the three specified ways, Holmes would satisfy as many instances of rationality requirements as he can and also violate as few of them as he can, given the situation. In case you find this recommendation utterly implausible, note that the formal model does *not* actually commit us to it. Equally well, it can deliver a stronger recommendation, that is, that *only one* of (i)–(iii) be followed. This is the topic of the next section.

Even though Holmes would end up violating a requirement no matter what he does, this doesn't mean that we should classify his response as irrational. Think back to the twins case. Suppose you wind up calling off your rendezvous with the first twin, Taylor, and dining with the second one, Tyler. There's something unfortunate

about your response—you have broken one of your promises—and yet it is an optimal response to the situation you were in. Similarly, we can suppose that Holmes complies with (iii) and ends up believing that the evidence doesn't support believing that the maid did it. There's something unfortunate about his response—he violates the evidential requirement by not having a belief supported by the evidence—and yet it is an optimal response to the situation he is in.

On the perspective that comes with the logic, requirements are defeasible. So, when assessing the agent's rationality, we should not be looking at whether or not she complies with all the requirements that are in force in her situation—formally, all the requirements in $Binding(c)$—but rather at whether or not the agent complies with the generated oughts. Thus, if any of (i)–(iii) obtain, Holmes' response is fully rational, and that's enough to conclude that Moriarty's Drug is not a genuine dilemma.

So we have a solution to the puzzle that concedes that the possibility of radically misleading total evidence (MTE) leads to a conflict between the epistemic requirement (ER) and the inter-level coherence requirement (ILC), and yet does not qualify it as a dilemma. Its advantages over the alternatives should be obvious: It lets us preserve (MTE), (ER), (ILC), as well as the unity of epistemic rationality.

The solution changes the logic governing the interaction between (ER) and (ILC), or, what's in effect the same thing, proposes that we think of them not as strict requirements specifying what doxastic attitudes one ought to have all things considered, but as defeasible requirements specifying what attitudes one should have ideally, requirements that interact to jointly determine the all things considered epis-

temic oughts—much like moral reasons are standardly taken to determine the all things considered moral oughts. I think it is very natural to call these defeasible requirements *ideals* and to think of the solution just developed as a mathematically precise characterization of the conflicting-ideals response to the puzzle. So my suggestion is that we haven't only solved the puzzle, but also taken some steps toward making the conflicting-ideals view itself more transparent. That is, I think, that it is best understood as a move away from the default metaepistemological opinion, according to which epistemic requirements are strict and governed by some rather strong, but never explicitly stated logic, toward the more unconventional metaepistemological view, according to which epistemic requirements are defeasible and governed by a comparatively weak logic. I don't think this interpretation of the view is standard: In the end, in the literature focusing specifically on the puzzle, Christensen (2007a, 2013) is usually interpreted as conceding the existence of genuine rational dilemmas.[30]

### 3.2.3 Disjunctive oughts and weighted requirements

Given the logic from Section 3.2.1, the context $c_3$ that we used to encode Moriarty's Drug entails a disjunctive ought, suggesting that Holmes can respond to this scenario in three different, but equally rational ways. One may find this worrisome on at least three grounds. First, one could feel that this diagnosis is simply intuitively implausible. Second, one could insist that this makes the logic a poor candidate for the formal backbone of the conflicting-ideals view, since the

[30]See footnote 8 for references.

latter isn't typically associated with disjunctive recommendations. And third, one could point out that this commits one to *permissivism*—or the view that there are at least some bodies of evidence sanctioning multiple rational responses—which we may have independent reasons to reject.[31] All of these worries cast doubt on the solution. The main goal of this section, then, is to put them to rest by showing that the solution is compatible with the claim that there's only one rational response to Moriarty's Drug.

Once we have stepped away from the assumption that *epistemic* rationality requirements have to be strict, it is natural to think that one requirement (or ideal) can have more weight than another, or that a particular instance of a requirement can have more weight than another. In particular, we might reasonably suspect that the requirements in force in Moriarty's Drug may have different (relative) weight. And if these weights are indeed different, then we would, of course, want them to be reflected in the context capturing the scenario. So it seems perfectly reasonable to insist that the context $c_3$ has to be supplied with a priority relation if the scenario is to be captured in full, and that, once it is thus extended, Holmes will have only one rational response.[32] The difficult question, then, is what should this ordering be.

Perhaps, one might think that any assignment of relatives weights to instances of epistemic requirements has to be uniform and motivated on independent grounds. For us here this would mean that the ordering would need to be supported by an argument for one of the following two claims: (ER) must always take precedence over

[31] For good discussions of permissivism and further pointers to the literature, see (White 2005) and (Schoenfield 2014).

[32] Generally, adding an ordering on requirements results in a stronger consequence set, and so an ordering on requirements in $c_3$ would lead to Holmes having fewer rational responses.

(ILC). Or (ILC) must always take precedence over (ER). However, both of these claims are in tension with the spirit of the conflicting-ideals view and our solution to the puzzle. To see why, suppose that we did hold that in every situation of conflict between the defeasible (ER) and the defeasible (ILC) the former wins out. Now it is very natural to ask what sort of work the defeasibility of (ER) is even doing and to wonder if the resulting position isn't better thought of as one that denies that (ILC) is a genuine requirement. So I think that, once we have embraced the idea that (ER) and (ILC) are ideals that occasionally come into conflict we can't uniformly prioritize one over the other.

What we can do, however, is hold that every particular conflict between instances of (ER) and (ILC) gets resolved, with (ER) winning out in some of them and (ILC) in others and with the winner being determined by the details of the case. Notice that this is just what we observe in cases of conflict between requirements in the moral domain. In the Drowning Child scenario it was clear that saving the child and not keeping the promise was the only right thing to do. But this doesn't mean that considerations of benevolence are always more important than those of promise-keeping. In fact, it's very easy to think of scenarios where keeping a promise and not doing the good seems like the only right thing to do. (Suppose a dog was drowning instead of a child; or that you knew that the child wasn't in real danger and would only calm down a bit if you helped it.) So sometimes benevolence wins out, and at other time promise-keeping does.[33] And it seems perfectly reasonable to

[33]Interestingly, Ross appears to have thought that promise-keeping is standardly more important than doing the good: "Ross [..] held that ordinarily considerations of promising should take precedence over those of beneficence (it being more important ordinarily to keep one's promise than to do good); but he also supposed that there could be such a thing as a trivial breach of promise and

hold that every particular conflict between benevolence and promise-keeping has a correct resolution.

But what are we to say about assigning weights to the requirements in $c_3$ in light of this? I think the correct answer here is that Moriarty's Drug is too underdescribed—and in a way that's typical of other examples illustrating (MTE) in the literature—for us to say what this ordering should be. We know that Holmes' first-order evidence, or the clues that he finds at the manor, supports believing that the maid did it, but we don't know what exactly this evidence is. Similarly, we know that Holmes knows how Moriarty's drug works, but we don't know what evidence exactly this knowledge is based on. A full description of Holmes' evidence would put us in a position to say what the weights of the relevant instances of (ER) and (ILC) are. But, in the absence of such a description, we can't (and shouldn't) say what they are or what exactly the one rational doxastic response to Moriarty's Drug is.

At this point one might object that, quite independently of the details suppressed in the description, it is actually intuitively very clear that Holmes' only rational response in Moriarty's Drug is to believe that his evidence doesn't support believing that the maid did it and not to believe that she did. But if there's such an intuition, it may stem from the implicit assumption that Holmes' first-order evidence has to be fairly complex and that only relatively elaborate reasoning can get one from it to the conclusion that the maid did it—in the end, that's what we typically see in murder mysteries. But suppose that the evidence pointing to the maid

a very large good to be done, in which case, he thought, the promise should be broken" (Dancy 2004, p. 27).

as the likely culprit was overwhelming: She has a clear motive, there are no other suspects, multiple witnesses report her having had obsessive thoughts of violence, and, on top of that, there's a video of the murder caught on a security camera with the murderer looking just like the maid. Further, suppose that Holmes' knowledge of the drug's effects was based on cases where the affected detectives reasoned about complex bodies of evidence. Would we still say that it would be irrational for Holmes to believe that the maid did? My own intuition says no.

Now let me close by discussing two sample assignments of weights to the requirements $r_4$, $r_5$, and $r_6$. The first would correspond to a fuller description we started sketching in the previous paragraph. Let $c_4 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a context that is like $c_3$, except for its $\leq$ is not empty, but as follows: $r_6 < r_5 < r_4$. Thus, both instances of (ER) have more weight than the instance of (ILC), and the instance of (ER) concerned with the first-order belief that the maid did it has more weight than the one concerned with the second-order belief.[34] A by now routine check will convince you that only the two strongest requirements $r_4$ and $r_5$ get selected as binding, and that the context $c_4$ entails the formula $\bigcirc[Believe(E \not\Mapsto Believe(M)) \& Believe(M)]$. So there's only one way for Holmes to be rational: He must believe that the maid did it, as well as believe that his evidence doesn't support believing that she did.

[34] What justifies ordering the instances of (ER)? Well, I think it's very plausible to think that, when we have both $E \Mapsto Believe(X)$ and $E \Mapsto Believe(Y)$, it doesn't yet mean that $E$ supports the beliefs $Believe(X)$ and $Believe(Y)$ equally well. Further, it seem reasonable to hold that the relative degree of support that the evidence lends to $Believe(X)$ and $Believe(Y)$ is what determines the relative weights of the corresponding requirements $\dfrac{E \Mapsto Believe(X)}{Believe(X)}$ and $\dfrac{E \Mapsto Believe(Y)}{Believe(Y)}$. In the particular case at hand, $Believe(M)$ appears to be supported better than $Believe(E \not\Mapsto Believe(M))$, and so I rank $r_4$ above $r_5$.

The second sample assignment would correspond to a different way of filling in the details of the scenario, one where the first-order evidence is indeed complex and calls for an elaborate chain of reasoning. Let $c_5 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be the context acquired by substituting the ordering $r_4 < r_6 < r_5$ for the empty ordering of $c_3$. Here the instance of (ILC) has more weight than one of the instances of (ER), but not the other. Yet again, only the two strongest requirements $r_5$ and $r_6$ get selected as binding. This time, however, the context entails a different formula, namely, $\bigcirc[Believe(E \not\Rightarrow Believe(M)) \& \neg Believe(M)]$. So, again, there's only one way for Holmes to respond rationally: He must have the second-order belief and avoid believing that the maid did it.

Note that $c_4$ and $c_5$ are only illustrations. I'm not suggesting that one of them reflects the correct assignment of weights (although it might). Instead, I'm suggesting the following: If we want to insist that Holmes must have only one rational response in Moriarty's Drug, we can reasonably hold that the relative weights of the requirements in play ensure that. So we are not committed to the disjunctive recommendation discussed at the end of Section 3.2.2. However, appealing to weighted requirements isn't strictly necessary, as we have a solution to the puzzle with or without it.

## 3.3 Summary

The goal of this chapter was to present a solution to an important puzzle that starts with acknowledging the possibility of total evidence that's radically misleading

about itself, (MTE), proceeds to a clash between two plausible and widely accepted requirements of rationality, (ER) and (ILC), and arrives at the claim that there are rational dilemmas. Aiming to avoid the conclusion, the existing responses to the puzzle have proposed rejecting (MTE), rejecting one of the requirements, or treating the conflict as a clash between two irreducible types of rationality. The main idea behind my alternative solution was to substitute a defeasible deontic logic for the relatively strong, but never explicitly stated logic governing the interaction between (ER) and (ILC). The advantages of the solution are obvious: It lets us preserve (MTE), (ER), (ILC), as well as the unity of epistemic rationality, all while steering clear of dilemmas. We saw that the use of the logic comes with an unorthodox perspective on epistemic normativity, namely, one on which what one epistemically ought to do is determined through the interaction of defeasible rationality requirements that apply to them. We also saw that such requirements can be understood as epistemic ideals, and that the logic together with the perspective it gives rise to fit nicely with the conflicting-ideals view. What's more, the logic would seem to make the view itself more transparent.

The work begun in this part of the dissertation could (and should) be taken in three directions in the future. First, the framework we have discussed—that is, the formal model and the new perspective on requirements—could be applied to many other puzzles in which epistemic requirements come into conflict. Second, it could be of use in the context of the meta-level debate about the existence of epistemic dilemmas.[35] And third, it should also be worthwhile to think about alternative ways

[35]See e.g., (Hughes 2017) and (Leonard 2020).

of making sense of defeasible rationality requirements, that is, not as ideals. In the end, the perspective that comes with the use of our formal model is general, and the conflicting-ideals view may be only one attractive way of filling in its details.

# Part III DEFEASIBLE REASONING POLICIES

# Chapter 4: Conciliationism and the problem of self-defeat

This third part of the dissertation discusses another application of logics for defeasible reasoning in epistemology, this time in the context of the debate about the epistemic significance of disagreement. More specifically, it presents a generalization of the model developed in the previous chapters and uses it to, first, get a better handle on the *conciliatory* or *conciliationist views* on disagreement and the logical structure of conciliatory reasoning such views advocate for, and, second, address a pressing challenge for these views.[1] On the view that comes with the model, conciliationism is to be thought of as a (second-order) defeasible reasoning policy.

Think of your favorite question in philosophy. You've likely pondered on it for a long time, and you must have an opinion on the matter. And odds are you know someone who's pondered on the question at least as much, whose credentials are as good as yours, but who holds the opposite opinion. That is, you're in disagreement with an *epistemic peer*.[2] Should this fact make you less confident that your take on the issue is correct? According to the conciliatory views on disagreement we'll be exploring in this and the next two chapters it should. This answer has much intuitive appeal: The matter is complex, you aren't infallible, and one straightforward explanation for the disagreement is that you've made a subtle mistake when reasoning.

[1] The most well-known advocates of conciliatory views include Christensen (2007b, 2011, 2016), Elga (2007), Feldman (2005, 2006, 2009), and Matheson (2015c).

[2] Although there are number of accounts of epistemic peers in the literature, central to all of them is the idea that your epistemic peer is your epistemic equal. Two kinds of equality are standardly emphasized, namely, equality of evidence and equality of the peer's capacity to process this evidence. In what follows, I will rely on an intuitive understanding of the notion. For more on it see, e.g., (Gelfert 2011), (Matheson 2015c, Ch. 2), (Matheson 2018), and (Mulligan 2015).

An equally good explanation, of course, is that your opponent has made a mistake. But given that there's no good reason to favor the latter, reducing confidence still seems appropriate.

In spite of their intuitive appeal, conciliatory views are said to run into problems when applied to themselves, or when attempting to answer the question of what should one do in the context of disagreeing about the correct way to respond to disagreement. The problems are most transparent and easiest to explain for the more extreme conciliatory views. According to such views, when you hold a well-reasoned belief that $X$ and an equally informed colleague disagrees with you about whether $X$, you should lower your confidence in $X$ dramatically, or—to state it in terms of categorical beliefs—you should abandon your belief and suspend judgment on the issue.

Let's imagine that you've reasoned your way toward such a view, and that you're the sort of person who acts on the views they hold. Imagine further that you have a well-reasoned opinion on some other complex issue, say, you believe that we have a libertarian free will. Now, to your dismay, you find yourself in a crossfire: Your friend metaphysician Milo thinks that there's no libertarian free will, while your friend epistemologist Evelyn thinks that one shouldn't abandon one's well-reasoned opinion when faced with a disagreeing peer. Call this scenario *Double Disagreement*. The question now is how should you adjust your beliefs. For starters, your conciliatory view appears to self-defeat, or call for abandoning itself. Just instantiate $X$ with it! There's disagreement of the right sort, and so you should abandon your view. And to make the matters worse, there's something in the vicinity

of inconsistency around the corner. What are you to do about your belief in free will? Since there's no antecedent reason to start by applying the conciliatory view to itself, you've two lines of argument supporting opposing conclusions: That you should drop your belief in the existence of free will and that it's not the case that you should. On the one hand, it'd seem that you should drop the belief, in light of your disagreement with Milo and your conciliatory view. On the other, there's the following line of argument too. Your conciliatory view self-defeats, and, once it does, your disagreement with Milo looses its epistemic significance. But if the disagreement isn't significant for you, then it's fine for you to keep your belief in the existence of free will. And if that's so, then it certainly can't be the case that you should drop your belief.[3]

Of course, this is sketchy and quick, but you must agree that disagreement about the correct way to disagree presents the proponents of strong conciliatory views with two challenges. The first is that there are *possible scenarios*—of which Double Disagreement is only one example—in which their views appear to issue inconsistent directives.[4] And the second becomes manifest once we realize that the epistemic circumstances that the agent finds herself in in our imagined scenario don't appear to be all that different from the epistemic circumstances that many advocates of (strong) conciliatory views *actually* find themselves in. In the end,

[3]It's worth pointing out that your situation might be even worse. If we suppose, as seems reasonable, that one shouldn't be dropping one's well-reasoned beliefs willy-nilly, then we can reason to the conclusion that you should abandon your belief in the existence of free will and that you should also keep it.

[4]The concern that conciliatory views, strong and weak, issue inconsistent directives or are inconsistent is discussed in (Christensen 2013), (Decker 2014), (Elga 2010), (Littlejohn 2013, 2019), (Matheson 2015a,c), and (Weatherson 2013).

Evelyn has many real-world counterparts, including such illustrious philosophers as Thomas Kelly (2005, 2010), Titelbaum (2015), and Ralph Wedgwood (2010). So if we think that you should give up your conciliatory view in response to the disagreement with Evelyn, we'd seem to have to say that many actual advocates of conciliationism are not being rational in holding onto their views.[5] The following strikes me as uncontroversial: If one is forced to admit that one's view can issue inconsistent recommendations or that one can't rationally hold one's view, one is in trouble.[6] So, the proponents of strong conciliatory views face these challenges. What's more, Christensen (2013), Elga (2010), and others have forcefully argued that issues stemming from disagreement about the correct way to disagree cause problems for all types of conciliatory views, whether very strong or more moderate.[7]

But we're not going to focus on moderate conciliatory views here. Instead, we will devise a formal model implementing a very strong version of the view and study the model's behavior in such cases as Double Disagreement.[8] This will let us

[5]This concern that a conciliatonist must abandon her views by her own lights is discussed in, e.g., (Decker 2014), (Kelly 2005), (Littlejohn 2013), and (Matheson 2015a,c).

[6]Cf. (Christensen 2013), (Decker 2014).

[7]According to more moderate conciliatory views, when you hold a well-reasoned belief that $X$ and find yourself in a disagreement of the right sort, you should lower your confidence in $X$ at least a little. But by how much exactly? Well, typically such views require that, in answering this question, you factor in your own competence in reasoning about $X$-like matters, as well as your colleague's, or, rather, your degree of confidence in these competences. But, then, it's easy enough to imagine a scenario prompting the self-defeat of even the more moderate views: Just suppose that you find yourself disagreeing over $X$ and that your confidence in your own competence in reasoning about $X$-like matters is extremely low, while your confidence in your colleague's competence in reasoning about $X$-like matters is extremely high. See, e.g., (Christensen 2013), (Decker 2014), and (Elga 2010), (Littlejohn 2013) for more one this. For completeness, I should also note that the literature talks about two more concerns associated with disagreeing about disagreement. According to the first, a conciliationist has to abandon her view when repeatedly disagreeing about disagreement with a stubborn opponent—see (Decker 2014), (Elga 2010), and (Weatherson 2013). And according to the second, a conciliationist can't maintain any stable view of the right way to respond to disagreement—see (Christensen 2013) and (Weatherson 2013). Both of these concerns, however, have to do with belief change over time, and both make a number of substantive assumptions about the way conciliatory views work. I won't have anything to say about these concerns here.

[8]In fact, the view we're going to explore comes close to the infamous *Equal Weights View*,

get a better understanding of the logical structure of conciliatory reasoning and its behavior in cases involving disagreement about the correct way to disagree, as well as, eventually, provide adequate responses to both concerns.

The remainder of this dissertation is structured as follows. In Section 4.1, we'll be concerned with developing the model. We'll embed the core idea behind conciliationism into a *defeasible reasoner*, or a logic with a consequence relation at its core. In Section 4.2 we'll be concerned with expressing Double Disagreement in the model. A formal version of the concern about conciliatory views issuing inconsistent directives will emerge in Section 4.2.2. We'll then address it, in two steps, in Chapters 5 and 6. The second concern—that one can't hold a conciliatory view rationally, given the current state of epistemic opinion—will resurface and get addressed in Chapter 6.

## 4.1 Model conciliatory reasoner

### 4.1.1 Basic defeasible reasoner

This section defines a defeasible reasoner. The particular reasoner we'll be working with is a form of *default logic*, and it will generalize the model we used in Part II.[9] The core idea behind default logic is to supplement the standard (classical) logic with a special set of rules representing defeasible generalizations, so as to be

according to which, in a case of disagreement about $X$, you are to give a peer's confidence in $X$ the same weight as your own—see, e.g., (Elga 2007).

[9]The original formulation of default logic is due to Reiter (1980), but the current presentation is based on the more user-friendly version of Horty (2012). It's a well-known fact that default logic is more general than the defeasible logic defined in terms of maximally consistent subsets and classical consequence from Section 3.2.1—see e.g., (Horty 1994).

able to derive a stronger set of conclusions from a given set of premises. As before, our background language will be the language of ordinary propositional logic with the usual connectives. As for the default rules, we are going to represent them in a way that's familiar from Chapter 1, that is, either in a tree-form or as pairs of formulas. Thus, where $X$ and $Y$ are arbitrary propositions, both $\frac{X}{Y}$ and $\langle X, Y\rangle$ will stand for the rule that lets us conclude $Y$ form $X$ by default. To take a simple example, let $B$ be the proposition that Tweety is a bird and $F$ the proposition that Tweety flies. Then $\frac{B}{F}$ says that we can conclude that Tweety flies as soon as we have established that he is a bird. We will use the letter $r$ to denote default rules, and make use of the familiar functions $Premise[\cdot]$ and $Conclusion[\cdot]$ to pick out, respectively, the premise and the conclusion of some given rule. If $r = \frac{X}{Y}$, then $Premise[r] = X$ and $Conclusion[r] = Y$. Yet again, we apply the second function not only to individual default rules, but also sets of rules: Where $\mathcal{S}$ is a set of rules, $Conclusion[\mathcal{S}] = \{Conclusion[r] : r \in \mathcal{S}\}$.

We envision an agent reasoning on the basis of a two-part structure $\langle \mathcal{W}, \mathcal{R}\rangle$ where $\mathcal{W}$ is a set of ordinary propositional formulas—the hard information, or the information that the agent is certain of—and $\mathcal{R}$ is a set of default rules—the rules the agent relies on in its reasoning. Just as we did in the previous chapters, we will call such structures *contexts* and denote them by the letter $c$, with subscripts.

**Definition 4.1 (Contexts)** *A* context *$c$ is a structure of the form $\langle \mathcal{W}, \mathcal{R}\rangle$, where $\mathcal{W}$ is a set of propositional formulas and $\mathcal{R}$ is a set of default rules.*

At this point an example is in order, and we'll use the one that's standard in the

artificial intelligence literature, the Tweety Triangle. It unfold in two steps. In the first, the reasoning agent learns that Tweety is a bird and infers that Tweety flies. In the second, it learns that Tweety is a penguin, retracts the previous conclusion, and infers that Tweety doesn't fly.

Since there are two steps, there'll be two contexts, $c_1 = \langle \mathcal{W}, \mathcal{R} \rangle$ and $c_2 = \langle \mathcal{W}', \mathcal{R} \rangle$. Let $B$ and $F$ be as before, and let $P$ stand for the proposition that Tweety is a penguin. The hard information $\mathcal{W}$ of $c_1$ must include $B$ and $P \supset B$, expressing an instance of the fact that all penguins are birds. The set of rules $\mathcal{R}$ of $c_1$, in turn, will contain the two rules $r_1 = \frac{B}{F}$ and $r_2 = \frac{P}{\neg F}$. The first lets the reasoner infer that Tweety can fly, by default, once it has concluded that Tweety is a bird. The second lets the reasoner infer that Tweety cannot fly, by default, once it has concluded that he is a penguin. Thus, $r_1$ and $r_2$ can be thought of as instances of, respectively, the idea that birds usually fly and the idea that penguins usually don't. As for $c_2 = \langle \mathcal{W}', \mathcal{R} \rangle$, it is just like $c_1$, except for its hard information also contains $P$, saying that Tweety is a penguin. We want the reasoner to concludes $F$ from $c_1$ and $\neg F$ from $c_2$. The question now is how can we make it do that.

It's worth saying upfront that the procedure for determining which formulas follow from a given context we are about to specify will be slightly more involved than the ones we have encountered in the previous chapters. Just like the others, it will rely on an intermediary notion, which will let us select all and only those rules of a given context that are *in force* or that *apply* in the case at hand. What will be different, however, is how this intermediary notion is defined.

We will start by introducing the concept of a *scenario* based on a context.

And a scenario based on a context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ is defined simply as some subset $\mathcal{S}$ of the set of rules $\mathcal{R}$ of this context. The intermediary notion selecting all the intuitively acceptable rules of $\mathcal{R}$ will be called a *proper scenario*, and our definition of a proper scenarios will emerge as a combination of three other notions, capturing the conditions on default rules from $\mathcal{R}$ that are necessary and jointly sufficient for a rule to be a part of a proper scenario. The first of these capture the familiar idea that a rule must be triggered.

**Definition 4.2 (Triggered rules, relativized)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ a scenario based on it. Then the default rules from $\mathcal{R}$ that are* triggered *in the scenario $\mathcal{S}$ are those that belong to the set $Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S}) = \{r \in \mathcal{R} : \mathcal{W} \cup Conclusion[S] \vdash Premise[r]\}$.*

Let's apply the function to the empty scenario $\varnothing$, against the background of the context $c_1$. It's easy to see that $Triggered_c(\varnothing) = \{r_1\}$, as the hard information $\mathcal{W} = \{B, P \supset B\}$ entails $B = Premise[r_1]$. You may wonder why we need to relativize the notion of a rule triggered in a context $c$ to that of a rule triggered in a scenario $\mathcal{S}$ based on a context $c$; and why I opted for $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Premise[r]$ in this definition, as opposed to the simpler $\mathcal{W} \vdash Premise[r]$.[10] The short answer is that we need it to handle situations where some default rules trigger further default rules. Examples of such situations will follow shortly.

The need for a further condition on proper scenarios reveals itself once we apply $Triggered(\cdot)$ to any scenario against the background of $c_2$. A minute's reflection

[10]Notice that in Chapters 1–3 we used the simpler expression.

reveals that both rules $r_1$ and $r_2$ come out triggered in every scenario based on this context. But the intuitively correct scenario based on it is $\{r_2\}$, and so we need to specify a further condition that would preclude the addition of $r_1$ to $\{r_2\}$. Here's a negative condition that does the trick:

**Definition 4.3 (Conflicted rules)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context, and $\mathcal{S}$ a scenario based on it. Then the rules from $\mathcal{R}$ that are* conflicted *in the context of $\mathcal{S}$ are those that belong to the set $Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S}) = \{r \in \mathcal{R} : \mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]\}$.*

Notice that we have $Conflicted_{\mathcal{W},\mathcal{R}}(\{r_2\}) = \{r_1\}$, as desired. Now consider the following preliminary definition for proper scenarios:

**Definition 4.4 (Proper scenarios, first pass)** *Let $\langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ a set of default rules. Then $\mathcal{S}$ is a proper scenario based on $\langle \mathcal{W}, \mathcal{R} \rangle$ just in case*

$$\mathcal{S} = \{r \in \mathcal{R} : \; r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S}), \; r \notin Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S})\}.$$

The definition gives us the correct result in the case of $c_1$. The singleton $\mathcal{S}_1 = \{r_1\}$ comes out as the only proper scenario. However, the definition falls flat when applied to $c_2$. Both $\mathcal{S}_1$ and $\mathcal{S}_2 = \{r_2\}$ qualify as proper. There are multiple ways to resolve the problem formally. The one I adopt here is motivated by the broader goal of this chapter (namely, that we need the resources that will let us model conciliatory views). We'll make use of an idea we briefly touched upon back in Section 2.2.2. Following Horty (2012), we will introduce a new type of rules, *exclusionary default*

*rules*, which you can think of as rules the reasoner uses to decide which other rules to take out of consideration.[11] In order to be in a position to formulate such rules, we extend the background language in two ways. First, we introduce names to refer to rules: Every default rule $r_X$ is assigned a unique name $\mathfrak{r}_X$—the Fraktur script is used to distinguish names of rules from the rules themselves. And second, we introduce a special predicate $Out(\cdot)$ into our language, with the intent that, where $\mathfrak{r}$ is a name of some default rule $r$, the statement $Out(\mathfrak{r})$ means that $r$ is excluded or taken out of consideration. For concreteness, we let $\mathfrak{r}_1$ be the name of $r_1$. Then $Out(\mathfrak{r}_1)$ says that $r_1$ is excluded.

With names and the new predicate in hand, we can state the second negative condition for a rule's inclusion in a proper scenario:

**Definition 4.5 (Excluded rules)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context, and $\mathcal{S}$ a scenario based on this context. Then the rules from $\mathcal{R}$ that are* excluded *in the context of $\mathcal{S}$ are those that belong to the set $Excluded_{\mathcal{W},\mathcal{R}}(\mathcal{S}) = \{r \in \mathcal{R} : \mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(\mathfrak{r})\}$.*

Our full definition of a proper scenario, then, runs thus:[12]

**Definition 4.6 (Proper scenarios)** *Let $\langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ a set of default*

[11] In Section 2.2.2, we briefly discussed the relation between exclusionary rules and hedged rules. In a nutshell, the idea was that we can model the same effects using both. The main reason I opt for using exclusionary rules in this part of the dissertation is that they are easier to work with and easier to visualize than hedged rules.

[12] I should point out that this definition ignores a technical problem that arises from the existence of aberrant contexts that contain self-triggering chains of rules. The simplest of such context is $\langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W} = \varnothing$ and $\mathcal{R} = \{ \frac{A}{A} \}$. But nothing important hinges on this. See (Horty 2012, p. 48f) for a discussion of the the problem and his Appendix A.1. for a solution.

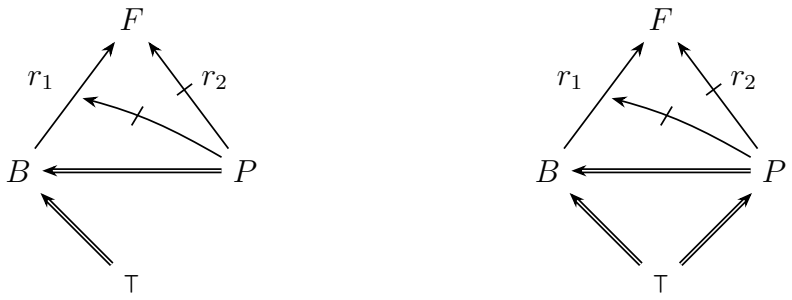



Figure 4.1: Tweety Triangle

*rules. Then* $\mathcal{S}$ *is a* proper scenario *based on* $\langle \mathcal{W}, \mathcal{R} \rangle$ *just in case*

$$\begin{aligned} \mathcal{S} \;=\; \{r \in \mathcal{R}: \;& r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S}), \\ & r \notin Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S}), \\ & r \notin Excluded_{\mathcal{W},\mathcal{R}}(\mathcal{S})\}. \end{aligned}$$

Returning to the example, $c_1$ and $c_2$ must be supplemented with the rule $r_3 = \dfrac{P}{Out(\mathfrak{r}_1)}$, saying that the rule $r_1$ must be taken out of consideration in case Tweety is a penguin. This should make good sense. Penguins form a rather specific type of birds. Once one learns that Tweety is a penguin, it wouldn't be wise to base the conclusion about his ability to fly on the idea that birds typically do. An easy check will convince you that $\mathcal{S}_1$ is still the only proper scenario base on $c_1$, and that $\mathcal{S}_3 = \{r_2, r_3\}$ is the only proper scenario based on $c_2$. So the addition of $r_3$ leaves us with a unique proper scenario in each case. But note that this won't hold in general, as there are contexts with multiple proper scenarios.

In what follows, I'll often represent contexts as *inference graphs* of the sort you can see in Figure 4.1, which depicts the Tweety Triangle. Here's how such graphs

should be red. A double link of the form $X \Longrightarrow Y$ stands for the proposition $X \supset Y$. As a special case, the link $\top \Longrightarrow Y$ stands for the idea that $Y$ is implied by an arbitrary tautology, or simply that $Y$ is true. A single link of the form $X \longrightarrow Y$ stands for a default rule of the form $\frac{X}{Y}$, while a crossed out link of the form $X \not\longrightarrow Y$ stands for a default rule of the form $\frac{X}{\neg Y}$. Finally, crossed out links that stars from a node $X$ and point to another link $r$ represent an exclusionary default of the form $\frac{X}{Out(\mathfrak{r})}$.

Now, with the notion of a proper scenario in hand, we can define a consequence relation, specifying what conclusions follow from any given context:[13]

**Definition 4.7 (Consequence)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context. Then the statement $X$ follows from $c$, written as $c \mathrel{|\!\sim} X$, just in case $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$ for each proper scenario $\mathcal{S}$ based on $c$.*

And with this, our basic defeasible reasoner is complete. We will interpret it as a model reasoner: If it outputs $X$ in the context $c$, then $X$ is the correct response to the situation that $c$ represents.

### 4.1.2 Capturing conciliationism

The next step is to see how we can use the reasoner to model conciliatory reasoning. Let's start by taking a close look at a well-known case, *Mental Math*, in which the conciliatory response seems particularly intuitive. It's nicely described in

[13]Note that this is only one of the consequence relations that can be defined in the present framework, the one that's usually called *skeptical*—see (Horty 2012, Sec. 1.3.1). Not much hinges on my choice of the consequence relation though: Almost all of the contexts we're going to discuss will have a single proper scenario, and, when there's a single proper scenario, alternative definitions of consequence produce the same results.

the following passage from (Christensen 2010a):

> **Mental Math**: My friend and I have been going out to dinner for many years. We always tip 20% and divide the bill equally, and we always do the math in our heads. We're quite accurate, but on those occasions where we've disagreed in the past, we've been right equally often. This evening seems typical, in that I don't feel unusually tired or alert, and neither my friend nor I have had more wine or coffee than usual. I get $43 in my mental calculation, and become quite confident of this answer. But then my friend says she got $45. I dramatically reduce my confidence that $43 is the right answer.[14]

Mental Math describes relatively complex reasoning, and we shouldn't miss the following three of its features. First, it consists of two distinct components, namely, the mathematical calculations and the reasoning prompted by the disagreement. Second, the agent's initial confidence in $43 being the correct answer is based entirely on her calculations, and is later reduced *because* the agent becomes suspicious of them.[15] And third, the disagreement is important precisely because it makes the agent suspect that she may have made a mistake in her calculations. It's legitimate to think of the mathematical calculations as the agent's first-order reasoning and her deciding how confident to be in the conclusions arrived at through these calculations as second-order reasoning.[16]

[14](Christensen 2010a, pp. 186–7).

[15]Notice that this makes sense only on the condition that an agent can take her mathematical reasoning to be fallible.

[16]It may also be worth pointing out hat conciliatory reasoning, in Mental Math and in other

Bearing this in mind, let's model the agent's reasoning in our formalism. (We'll be making a couple of stabs, gradually refining the modeling.) The first thing we need to do is introduce a new predicate to our language: $Seems(\cdot)$. Now, $Seems(X)$ expresses the idea that the agent has arrived at the conclusion $X$ through her first-order reasoning about whether $X$. Admittedly, I don't have too much to say about the reasoning presupposed by $Seems(\cdot)$, but it should be clear that it will depend on $X$. If $X$ is a mathematical proposition, $Seems(X)$ is a result of mathematical calculations of the sort we just saw. If $X$ is a philosophical claim, $Seems(X)$ is a result of a careful philosophical investigation. Either way, the idea is that the agent has reasoned to the best of her ability about some nontrivial matter and arrived at $X$ as a result. Note also that $Seems(X)$ is compatible with $\neg X$. A fallible rational agent knows that even her best reasoning doesn't guarantee that the conclusion is correct.

Still, situations in which an agent's best reasoning leads her astray will presumably be rare, and the agent will typically go by her best first-order reasoning. This motivates the following default rule schema:[17]

**Significance of first-order reasoning**:

$r(X) = \dfrac{Seems(X)}{X}$ : If your best first-order reasoning outputs $X$, conclude $X$ by default.

In Mental Math, the schema would get instantiated by the rule $r_4 = \dfrac{Seems(S)}{S}$

cases, appears most intuitive when we think of the case at hand from a first-personal perspective. This perspective will retain its significance in our formal model.

[17]Pollock (2008) discusses a similar schema, albeit in a different context.

, where $S$ stands for the proposition that my *s*hare of the bill is \$43. Now note that what the friend's announcement brings into question is exactly the connection between $Seems(S)$ and $S$. My first-order mathematical calculations are usually reliable, but I do make mistakes in some cases. The announcement, then, suggests that this may be one them.

In order to model the effects of the announcement, we will make use of a second predicate, $Disagree(\cdot)$. Let $Disagree(X)$ express the idea that the agent is in genuine disagreement about whether $X$. And I say *genuine* to distinguish the disagreements that provide the agent with a compelling reason to suspect that something may have gone wrong with her reasoning from what we might call *merely apparent disagreements*, examples of which would include verbal disagreements and disagreements with what the literature calls *epistemic inferiors* (people who are clearly incompetent to form well-reasoned beliefs on the question the disagreement is about). So $Disagree(S)$ means that there's genuine disagreement about whether my share of the bill is indeed \$43. As our first pass, we will try to capture the effects of the friend's announcement by means of the material implication $Disagree(S) \supset Out(\mathfrak{r}_4)$. It says that, if there is genuine disagreement about whether $S$, then the rule $r_4$ is to be taken out of consideration. Accordingly, the core of conciliationism would be captured by the following schema:

**Conciliationist reasoning policy, first pass**:

$r^*(X) = Disagree(X) \supset Out(r(X))$: If there's genuine disagreement about whether $X$, stop relying on your first-order reasoning about $X$.

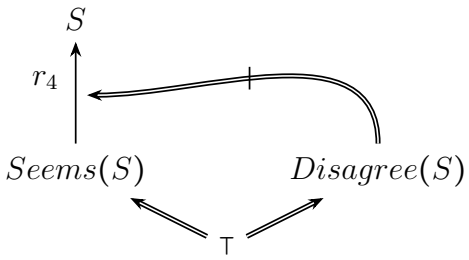



Figure 4.2: Mental Math, preliminary

We can try capturing Mental Math using the context $c_3 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W} = \{Seems(S),\ Disagree(S),\ Disagree(S) \supset Out(\mathfrak{r}_4)\}$ and $\mathcal{R} = \{r_4\}$. (The context is depicted in Figure 4.2.) An easy check will convince you that $S$ doesn't follow from $c_3$, as desired.

But $c_3$ doesn't capture all the complexities of Mental Math, as becomes clear once we consider the following question. Does the agent know for a fact that there's genuine disagreement about the bill, or does she only surmise it? I think it's obvious that it is the latter. Admittedly, the case doesn't describe the agent's reasoning to the existence of disagreement explicitly, but it is implicit in the background story: They've been going out to dinner for years, their track record is equally good, and so on. We may think that there's an easy fix, and that all we need to do is substitute the conciliationist schema for:

**Conciliationist reasoning policy, second pass**:

$r^{**}(X) = Seems(Disagree(X)) \supset Out(r(X))$: If your first-order reasoning about the presence of disagreement about $X$ suggests that there's genuine disagreement, stop relying on your first-order reasoning about $X$.

While this is a step forward, it turns out to be insufficient. To see why, let's consider a case discussed in the following passage from (Mulligan 2015):

> **Second-Order Disagreement**: Imagine that I disagree with my friend Francis about the truth of some proposition $Q$. I believe that $Q$ is true and Francis believes that $Q$ is false. Since I regard Francis as my epistemic peer with respect to $Q$, I revise my confidence in $Q$ downward... I subsequently hear something distressing from another friend, Richard, though: He believes that I erred when I judged Francis to be my epistemic peer. In Richard's opinion, Francis is not my epistemic peer. This is problematic because I take [it] that Richard is my epistemic peer with respect to assessments of epistemic peerhood.[18]

So here we have a case of genuine disagreement about whether there is disagreement about whether $Q$, or a case of second-order disagreement. Let's unpack the protagonist's reasoning. She starts off thinking that there's genuine disagreement about whether $Q$ between her and Francis, that is, the sort of disagreement that can only obtain between epistemic peers. Then she learns that Richard disagrees with her regarding Francis' status. If Richard turned out to be right, Francis wouldn't be her epistemic peer, which would, in turn, mean that her first-order reasoning about the existence of a genuine disagreement about whether $Q$ rested on a mistake. What's crucially important for us to realize is that the protagonist's disagreement with

[18](Mulligan 2015, p. 69), I've changed the propositional letter from $P$ to $Q$. Mulligan uses this case to exhibit two paradoxes in conciliatory views, and then goes on to discuss another similar case to exhibit a third one. The formal model developed here could help resolve these paradoxes, but we won't discuss the issue for reasons of space.

Richard doesn't show conclusively that her reasoning about whether there's genuine disagreement about whether $Q$ rests on a mistake. Rather, it only indicates it is a real possibility. In fact, here things aren't any different from any old case of disagreement: The protagonist has no way of knowing if she has made a mistake or not. Still, the disagreement with Richard is enough to make the first-order reasoning about the existence of genuine disagreement about whether $Q$ suspect. Thus, we have a scenario in which the agent's best first-order reasoning suggests that there's genuine disagreement about $Q$, that is, $Seems(Disagree(Q))$ obtains, and yet she can't justifiably reach the conclusion that there is genuine disagreement about $Q$, that is, she can't conclude $Disagree(Q)$.[19] But if that's so, then the above conciliatory schema $Seems(Disagree(X)) \supset Out(r(X))$ is much too strong. With it, the reasoner would proceed from $Seems(Disagree(Q))$ to disregarding its first-order reasoning about $Q$, but we have just seen that it shouldn't even proceed from $Seems(Disagree(Q))$ to $Disagree(Q)$.

Here we need to move from a strict connection between $Seems(Disagree(X))$ and $Out(\mathfrak{r}(X))$ to a defeasible one. And we're going to implement this idea by replacing the schema $Seems(Disagree(X)) \supset Out(r(X))$ with the two-link chain of default rule schemas, namely, $\frac{Seems(Disagree(X))}{Disagree(X)}$ and $\frac{Disagree(X)}{Out(r(X))}$. The first link makes the reasoner conclude, by default, that there's genuine disagreement about whether $X$ whenever its first-order reasoning suggests that there's genuine disagreement about whether $X$. Notice that this is nothing but the by now familiar

[19] By the way, I do think that there can just as well be cases of even higher-order disagreement, third-, forth-, and so on.

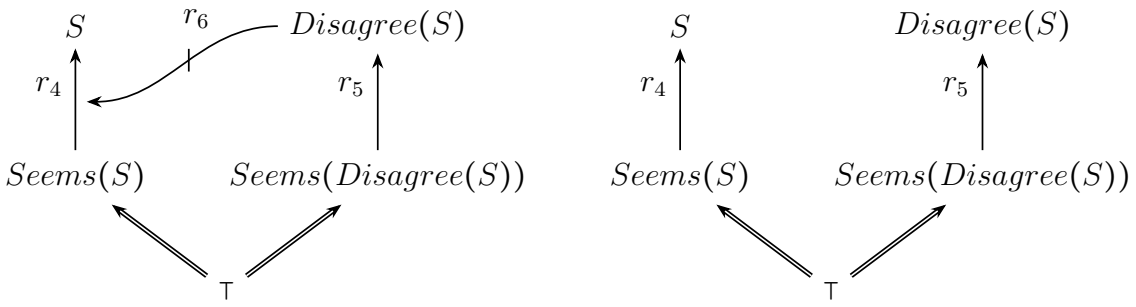



Figure 4.3: Mental Math, final

first-order reasoning rule schema $r(Y) = \frac{Seems(Y)}{Y}$ restricted to cases where this reasoning concerns disagreement. And the second link makes the reasoner take the rule of the form $\frac{Seems(X)}{X}$ out of consideration, again, by default, once the reasoner concludes that there's genuine disagreement about whether $X$.[20] I think that exactly this second link should be seen as heart of conciliatory views we are modeling.

**Conciliationist reasoning policy, final**

$r'(X) = \frac{Disagree(X)}{Out(r(X))}$ : If you have concluded that there's genuine disagreement about whether $X$, stop relying on your first-order reasoning about $X$ by default.

With this our conciliatory reasoner is complete, or nearly complete—see the following section. It's worth taking a look at our final formalization of the paradigm case of peer disagreement, Mental Math. We will now express it in the context $c_4$ =

[20]This is not the only way to implement the idea, but it is the most perspicuous one. Alternative implementations include changing the schema for $\frac{Seems(Disagree(X))}{Out(r(X))}$ and splitting it into $\frac{Seems(Disagree(X))}{Disagree(X)}$ and $Disagree(X) \supset Out(r(X))$. All three ways of implementing the idea lead to the similar analyses.

$\langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W}$ consisting of $Seems(S)$ and $Seems(Disagree(S))$ and $\mathcal{R}$ consisting of the default rules:

$$r_4 = \frac{Seems(S)}{S},$$

$$r_5 = \frac{Seems(Disagree(S))}{Disagree(S)}, \text{ and}$$

$$r_6 = \frac{Disagree(S)}{Out(\mathfrak{r}_4)}.$$

The first rule is familiar, but the remaining two are new. The rule $r_5$ makes the reasoner conclude that there's genuine disagreement about whether $S$, by default, if its first-order reasoning suggests that there's such disagreement. And the rule $r_6$ makes the reasoner take $r_4$ out of consideration, by default, once it concludes there's genuine disagreement about whether $S$. Figure 4.3 (left) depicts the context graphically. An easy check will convince you that $S$ does not follow from $c_4$, or $c_4 \not\vdash S$, as desired.

In the literature on peer disagreement, the views that are contrary to conciliatory views are called *steadfast*. So you may naturally wonder how such views would be represented in our model. My proposal is to represent them using a *steadfast reasoner*, or a reasoner that never makes use of the distinctively conciliatory schema $r'(X) = \frac{Disagree(X)}{Out(\mathfrak{r}(X))}$. How would we model a steadfast response in Mental Math then? Well, we would encode the scenario into the context $c_5 = \langle \mathcal{W}, \mathcal{R} \rangle$ which is just like $c_4$, except for the conciliatory rule $r_6$ is absent from it. Figure 4.3 (right) depicts this context graphically. It's easy to see that $S$ follows from $c_5$, as does $Disagree(S)$. So the steadfast reasoner concludes that there's genuine disagreement about whether its share of the bill is \$43, and concludes that it is \$43 anyway.

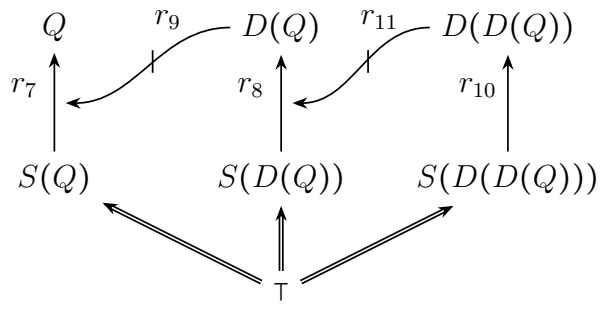



Figure 4.4: Second-Order Disagreement

We still need to capture the more complex case of Second-Order Disagreement. Before we do it, however, let's introduce some abbreviations. Henceforth, in graphs $Seems(\cdot)$ will sometimes be abbreviated as $S(\cdot)$ and $Disagree(\cdot)$ as $D(\cdot)$. We will encode the case itself in the context $c_6 = \langle \mathcal{W}, \mathcal{R} \rangle$. Its hard information $\mathcal{W}$ will consist of $Seems(Q)$, $Seems(Disagree(Q))$, and $Seems(Disagree(Disagree(Q)))$, with the last one expressing the proposition that the agent's first-order reasoning suggests that there's genuine disagreement about whether there is genuine disagreement about whether $Q$. The contexts set of rules, in turn, will include the following five:

$$r_7 = \frac{Seems(Q)}{Q},$$

$$r_8 = \frac{Seems(Disagree(Q))}{Disagree(Q)},$$

$$r_9 = \frac{Disagree(Q)}{Out(\mathfrak{r}_7)},$$

$$r_{10} = \frac{Seems(Disagree(Disagree(Q)))}{Disagree(Disagree(Q))}, \text{ and}$$

$$r_{11} = \frac{Disagree(Disagree(Q))}{Out(\mathfrak{r}_8)}.$$

The meanings of these default rules are straightforward to understand, and Figure 4.4 contains the graph depicting $c_6$. What conclusions does the reasoner draw from

this context? Well, here we have

$$c_6 \sim Q, Disagree(Disagree(Q)) \text{ and } c_6 \not\sim Disagree(Q).$$

So the reasoner concludes $Q$, as well as that there's genuine disagreement about whether there's genuine disagreement about whether $Q$, and it does *not* conclude that there's genuine disagreement about whether $Q$. What happens here is that the second-order disagreement cancels the effects of the first-order disagreement and reestablishes the link between $Seems(Q)$ and $Q$. This phenomenon is known as *reinstatement* in the logic literature: A rule $r$ gets excluded, but, then, through further instances of exclusion, $r$ reemerges, or gets reinstated, to support its original conclusion.[21] Thus, on our analysis, the agent should take her disagreement with Richard to nullify the effects of her disagreement with Francis, and she should go by her first-order reasoning about whether $Q$. I think this makes good sense. Given that we're in categorical-belief setting, the only other sensible thing to say here would be that the agent should suspend judgment about whether $Q$. But given that her best (first-order) reasoning points to $Q$, she would seem to need a compelling reason to suspend. Admittedly, a genuine disagreement regarding $Q$ could serve as such a reason, but then she has a compelling reason to suspend judgment on whether there is genuine disagreement regarding $Q$.

Before we leave this section and return to disagreement about the correct way to disagree, it's worth taking a brief look at a case of the sort that are often brought up against conciliationism to see how our model might handle it. Consider

[21] For more on reinstatement, see e.g., (Horty 2012, Section 8.3.3) and the references provided there.

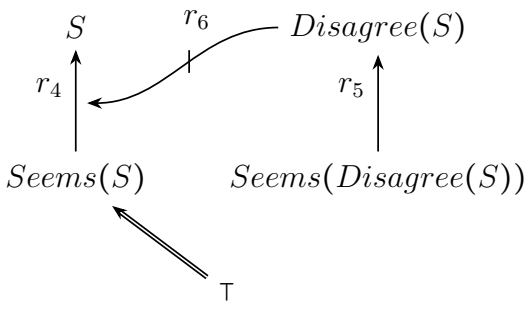



Figure 4.5: Careful Checking

the following variation on Mental Math:

> **Careful Checking**: I consider my friend my peer on matters of simple math. She and I are in a restaurant, figuring our shares of the bill plus 20% tip, rounded up to the nearest dollar. The total on the bill is clearly visible in unambiguous numbers. Instead of doing the math once in my head, I take out a pencil and paper and carefully go through the problem. I then carefully check my answer, and it checks out. I then take out my well-tested calculator, and redo the problem and check the result in a few different ways. As I do all of this I feel fully clear and alert. Each time I do the problem, I get the exact same answer, $43, and each time I check this answer, it checks out correctly. Since the math problem is so easy, and I've calculated and checked my answer so carefully in several independent ways, I now have an extremely high degree of rational confidence that our shares are $43. Then something very strange happens. My friend announces that she got $45 (Christensen 2011, p. 8).

In Careful Checking the intuition suggests, seemingly contra conciliationism, that

the agent should not reduce her confidence in $S$. It's very unlikely that someone in her situation has gotten the same wrong answer each time, and I think Christensen (2011) is absolutely right when he says that here the agent has good reason to suspect that something screwy is going on with her friend. So Careful Checking is not only a situation in which there is *no* genuine disagreement about whether $S$, but also a situation in which the agent doesn't arrive at the conclusion that there's such disagreement through her first-order reasoning. Or, to put it simply, $Seems(Disagree(S))$ doesn't obtain. We can encode the case in the context $c_7 = \langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W} = \{Seems(S)\}$ and $\mathcal{R} = \{r_4, r_5, r_6\}$, that is, a context that's just like $c_4$ which we used to encode Mental Math, except for now the proposition $Seems(Disagree(S))$ is absent from the hard information. Figure 4.5 depicts $c_7$ graphically. Not surprisingly, we get the intuitive result $c_7 \mathrel{|\!\sim} S$.[22]

## 4.2 Disagreement about conciliationism

Having embedded conciliationism in our defeasible reasoner, we can return to the problems stemming from disagreement about the correct way to disagree. Recall the problematic scenario that we started with. For convenience, let's restate it here,

[22]Some might find this way of dealing with the case dissatisfying. They might ask, "If, for any case at hand, we're allowed to decide whether the agent's first-order reasoning suggests that there's genuine disagreement or not, don't we have a response available to every case that causes trouble for conciliationism?" I have two things to say in response to this question. First, this move isn't going to work for all seemingly problematic cases—for instance, it's not going to help with Double Disagreement. Second, in Chapter 6 we are going to implement the idea of degrees of confidence in the model. This will let us capture formally such claims as, "The agent is more confident in the conclusion of her-first order reasoning that her share of the bill is \$43 than in the conclusion of her-first order reasoning that there's genuine disagreement about whether her share of the bill is \$43." With these degrees in hand, we could represent the scenario more adequately, but the reasoner would still draw the same conclusion.

in first-personal terms:

> **Double Disagreement**: I consider myself a somewhat able philosopher with special interests in metaphysics and social epistemology. I've reasoned very carefully about the vexed topic of free will, and I've come to the conclusion that we have libertarian free will. I've also spent a fair amount of time thinking about the issues surrounding peer disagreement, and as a result I've become convinced that conciliationism is correct and that on has to give up one's well-reasoned opinion when faced with a disagreeing peer. Then, to my amazement, I discover that my friend metaphysician Milo disagrees with me about the existence of free will. What's more, my friend epistemologist Evelyn disagrees with me about conciliationism. In fact, she thinks it's utterly misguided.[23]

Elga (2010), among others, has famously argued that cases like this show that conciliatory views are inconsistent.[24] His line of reasoning goes roughly as follows. On the one hand, conciliationism seems to recommend that the agent abandons her belief in the existence of libertarian free will and suspend judgment on the issue in response to her disagreement with Milo. On the other hand, conciliationism also seems to recommend that the agent does *not* abandon her belief in the existence

[23]This scenario is a variation on a case discussed by Matheson (2015a), see also (Christensen 2013).

[24]To be precise, Elga doesn't discuss any case like Double Disagreement. Instead, he draws an analogy between conciliatory views and the magazine *Consumer Reports* that reviews products, as well as other consumer ratings magazines, and ends up giving inconsistent advice: To buy only toaster $X$ and to follow the advice of another magazine, *Smart Shopper*, that suggest buying only toaster $Y$. However, the situation with magazines is supposed to be structurally analogous to some case with disagreement about conciliationism, and that case must look something like Double Disagreement.

of libertarian free will. Why? Well, it recommends taking her disagreement with Evelyn seriously. And once she does take it seriously, she should give up her belief in conciliationism as the correct response to cases of disagreement. But once the agent does that, her disagreement with Milo looses its epistemic significance for her. And if this disagreement has lost its epistemic significance, then it must be okay for the agent to retain her belief in the existence of free will. So it can't be the case that she has to abandon this belief. Thus, conciliationism, appears to support two inconsistent conclusions, that the agent should suspend belief with regard to whether free will exists and that it's not the case that the agent should suspend belief with regard to whether free will exists. From here, conciliationism is inconsistent, and so it must be abandonend, or, at least, substantially modified.

While the above reasoning has some pull, it's hard to deny that it could be a little more precise. So let's see if a formal analysis drawing on our model will support it. The first thing we need to do is encode the case in a context, and here already we face a difficulty. Part of the agent's reasoning concerns conciliationism itself. To be clear, we know how to model conciliatory reasoning, but we don't know yet how to model the reasoning that would put conciliationism as a reasoning policy into place. As our first step, we'll try treating conciliationism as an atomic proposition. So let $C$ stand for the idea that conciliationism is correct—think of $C$ as placeholder to be made precise later on—and let $L$ stand for the idea that we have libertarian free will. Now we can encode the scenario in the context $c_8 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ is composed of $Seems(L)$, $Seems(C)$, $Seems(Disagree(L))$, and $Seems(Disagree(C))$, and where $\mathcal{R}$ contains the following rules:

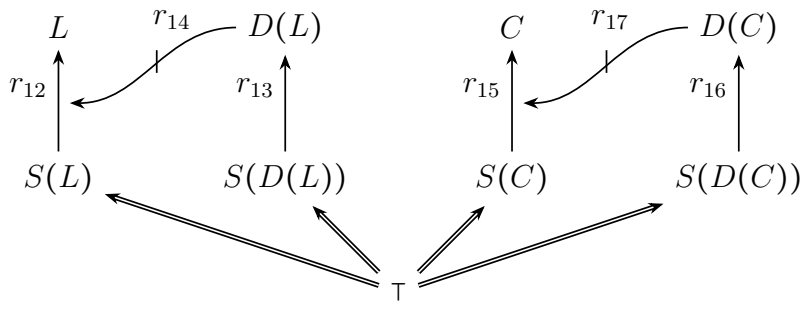



Figure 4.6: Double Disagreement, preliminary

$r_{12} = \dfrac{Seems(L)}{L}$ : If my best reasoning about free will suggest that $L$, I can conclude that $L$ by default.

$r_{13} = \dfrac{Seems(Disagree(L))}{Disagree(L)}$ : If my best reasoning about the disagreement between Milo and me suggests that there's genuine disagreement regarding $L$, I can conclude that there's genuine disagreement regarding $L$ by default.

$r_{14} = \dfrac{Disagree(L)}{Out(\mathfrak{r}_{12})}$ : If there's genuine disagreement regarding $L$, I must back off from my first-order reasoning about $L$ by default.

$r_{15} = \dfrac{Seems(C)}{C}$ : If my best reasoning about the epistemic significance of peer disagreement suggests that $C$, I can conclude that $C$ by default.

$r_{16} = \dfrac{Seems(Disagree(C))}{Disagree(C)}$ : If my best reasoning about the disagreement between Evelyn and me suggests that there's genuine disagreement regarding $C$, I can conclude that there's genuine disagreement regarding $C$ by default.

$r_{17} = \dfrac{Disagree(C)}{Out(\mathfrak{r}_{15})}$ : If there's genuine disagreement regarding $C$, I must back off from my first-order reasoning about $C$ by default.

The context $c_8$ is depicted in Figure 4.6. There's only one proper scenario based on

$c_8$, namely, $\mathcal{S}_4 = \{r_{13}, r_{14}, r_{14}, r_{17}\}$. From here it's but one step to see that $c_8 \not\vdash L$ and $c_8 \not\vdash C$. So the model suggests that the correct conciliatory response in Double Disagreement is to withhold belief on the question about the existence of libertarian free will, as well as the the question about the correctness of conciliationism. There is nothing blatantly inconsistent here, but the suggested responses is at least somewhat odd: Even though the reasoner withholds with regard to $C$, the derivation of $L$ is blocked for a distinctively conciliatory reason—the rule $r_{12}$ is defeated by the rule $r_{14}$. This, however, shouldn't be all that surprising. For, in $c_8$, there's no connection between reasoning to the conciliatory reasoning policy and this policy itself, or the rules implementing this policy. What our formalization is lacking is a connection between $C$ and the rules $r_{14}$ and $r_{17}$.

We're now going to put this connection into place, proceeding in two steps. In Section 4.2.1 we'll be concerned with the relation between $C$ and $r_{14}$ and the question of what happens if the support for $C$ gets undermined. And in Section 4.2.2 we'll turn to the connection between $C$ and $r_{17}$.

### 4.2.1 Reasoning to the conciliatory policy

To isolate the first task, let's refocus on a different case, one that's much like Double Disagreement, except for now, instead of discovering that I'm in disagreement with Evelyn, I find out that I've been given one of the infamous reasoning-distorting drugs:

**Disagreement on Drugs**: I consider myself a somewhat able philoso-

> pher with special interests in metaphysics and social epistemology. I've reasoned very carefully about the vexed topic of free will, and I've come to the conclusion that we have libertarian free will. I've also spent a fair amount of time thinking about the issues surrounding peer disagreement, and as a result I've become convinced that conciliationism is correct and that on has to give up one's well-reasoned opinion when faced with a disagreeing peer. Then I discover that my friend metaphysician Milo disagrees with me about the existence of free will. What's more, I find out, from a very reliable source, that someone has slipped a drug into my morning coffee, a drug that's known to screw up one's reasoning in matters pertaining to issues in epistemology, while leaving one's reasoning about matters pertaining to metaphysics intact.

Let $Drug$ express the idea that I've been given the drug, having the described effects. Now we can encode the scenario into $c_9 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W} = \{Seems(L), Seems(Disagree(L)), Seems(C), Seems(Drug)\}$ and $\mathcal{R}$ contains the familiar rules $r_{12}$, $r_{13}$, $r_{14}$, $r_{15}$ together with the new rule $r_{18} = \dfrac{Seems(Drug)}{Out(\mathfrak{r}_{15})}$. The new element $Seems(D)$ expresses the idea that I've arrived at the conclusions that I've been given the drug through appropriate first-order reasoning; and the new rule $r_{18}$ says if that's what my first-reasoning suggests, then I'd better stop relying on the reasoning affected by this drug.[25] The context is depicted in Figure 4.7. Note, though, that $c_8$ is only a preliminary formalization.

[25]Note that we could split $r_{18}$ into two default rules, with one taking the reasoner from $Seems(Drug)$ to $Drug$ and the other from $Drug$ to $Out(\mathfrak{r}_{15})$. Nothing of importance would change had we opted for this alternative representation.

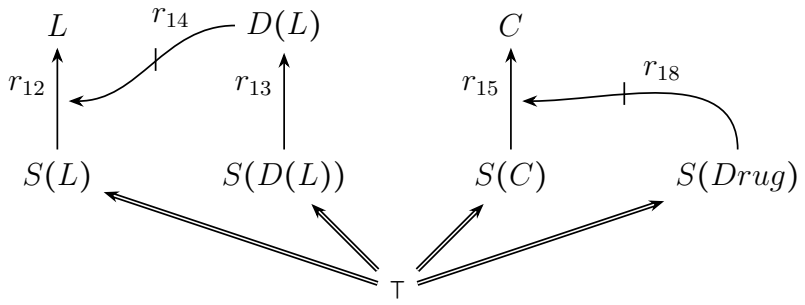



Figure 4.7: Disagreement on Drugs, preliminary

If we think about the case intuitively, we'd seem to have to say that the agent's conclusion that conciliationism is correct is what makes her adopt the conciliatory reasoning policy. In the model, the conclusion is represented by the propositional letter $C$ and the policy is represented as the rule schema $r'(X) = \dfrac{Disagree(X)}{Out(\mathfrak{r}(X))}$, where $r(X) = \dfrac{Seems(X)}{X}$. And in the particular case of $c_8$ the schema has only one instance, namely, the rule $r_{14} = \dfrac{Disagree(L)}{Out(\mathfrak{r}_{12})}$. So it seems reasonable to arrange things in such a way that $C$ is just what makes the reasoner adopt the rule $r_{14}$, just what puts this rule into place.

As the first step toward enabling the reasoner to adopt new rules, we'll extend the background language with a new predicate $Reasonable(\cdot)$. Think of it as the opposite of the predicate $Out(\cdot)$. If $Out(\mathfrak{r})$ says that the rule $r$ should be taken out of consideration, then $Reasonable(\mathfrak{r})$ says that $r$ is a reasonable rule to follow, or that $r$ should be among the reasoner's stock of rules for reasoning. The expression $Reasonable(\mathfrak{r}_{14})$ then expresses the idea that the rule $r_{14}$ in particular is a reasonable rule to follow.

We'll need to update the reasoner so that it can reason with the new predicate. But first let's think of a good way to ensure that the reasoner concludes

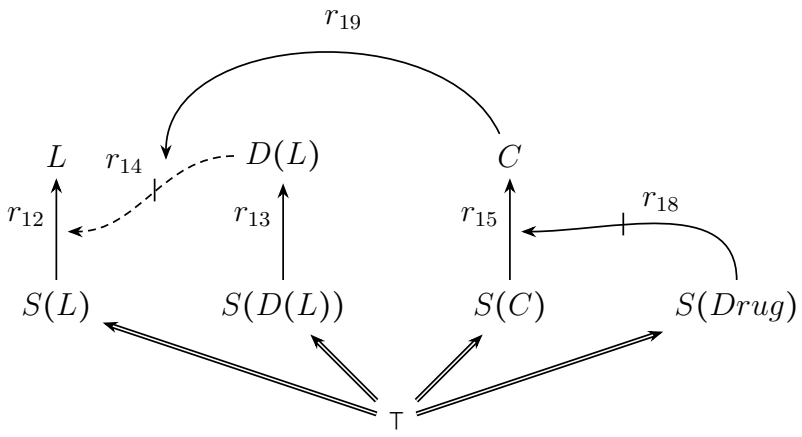



Figure 4.8: Disagreement on Drugs, final

$Reasonable(\mathfrak{r}_{14})$ in the context $c_8$. One straightforward thing we can do is add the rule $r_{19} = \dfrac{C}{Reasonable(\mathfrak{r}_{14})}$ to the context's set of rules. Notice that it lets the reasoner conclude, by default, that $r_{14}$ is a reasonable rule to follow, once it has concluded that conciliationism is correct. This results in the context $c_9 = \langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W} = \{Seems(L), Seems(Disagree(L)), Seems(C), Seems(Drug)\}$ and $\mathcal{R} = \{r_{12}, r_{13}, r_{14}, r_{15}, r_{18}, r_{19}\}$. See Figure 4.8 for the corresponding graph. Notice that the graph contains a dashed link. Henceforth, we will use such links to represent those rules that the reasoner can start relying on only after it has concluded that they are reasonable rules to follow. Links that point to dashed links represent rules of the form $\dfrac{X}{Reasonable(\mathfrak{r})}$.

If we run the reasoner on $c_9$, we see that neither $L$, or $Reasonable(\mathfrak{r}_{14})$ follow from it. But this is, of course, the wrong result. For, first, intuitively, in Disagreement on Drugs the agent should conclude that we have libertarian free will—or, at least, that's something we might argue for. Second and more importantly for our present concerns, $L$ doesn't follow because it is excluded by $d_{14}$, and so the reasoner follows

a rule, even though it hasn't deemed this rule a reasonable one to follow.

In order to make the new predicate and rules like $r_{19}$ do real work, we have to change the inner workings of the defeasible reasoner. And here my general strategy is to let the reasoner employ *any* default rule $r$ only on the condition that it can also infer a proposition of the form $\mathit{Reasonable}(\mathfrak{r})$, where $\mathfrak{r}$ is the unique name of $r$. This will run analogously to the way it works for a rule's triggering conditions. Currently, the reasoner can use a rule $r$ only in case it can infer its triggering condition, $\mathit{Premise}[r]$. Henceforth, however, it will use a rule $r$ only in case it can infer both $\mathit{Premise}[r]$ and $\mathit{Reasonable}(\mathfrak{r})$. And just like it worked for triggering conditions, there will be two ways for the reasoner to infer a formula of the form $\mathit{Reasonable}(\mathfrak{r})$, either from the hard information $\mathcal{W}$, or by means of other rules. One implication of this is that we will often have to include such formulas in $\mathcal{W}$. But this should make good sense: The presence of $\mathit{Reasonable}(\mathfrak{r})$ in $\mathcal{W}$ can be understood in terms of the reasoner being committed to $r$ from the outset of its reasoning.

Recall that, in Section 4.1.1, we defined the central notion of a proper scenario by specifying three conditions for a rule's inclusion in such a scenario. We will now amend this definition by adding a fourth condition—a condition requiring that the default rule is deemed reasonable:

**Definition 4.8 (Reasonable rules)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context, and $\mathcal{S}$ a scenario based on it. Then the rules from $\mathcal{R}$ that are* reasonable *(to follow) in the context of the scenario $\mathcal{S}$ are those that belong to the set $\mathit{Reasonable}_{\mathcal{W},\mathcal{R}}(\mathcal{S}) = \{r \in \mathcal{R} : \mathcal{W} \cup \mathit{Conclusion}[S] \vdash \mathit{Reasonable}(\mathfrak{r})\}$.*

With this definition in hand, we can update the definition of a proper scenario, as follows:

**Definition 4.9 (Proper scenarios, updated)** *Let $\langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ a scenario based on it. Then $\mathcal{S}$ is a* proper scenario *based on $\langle \mathcal{W}, \mathcal{R} \rangle$ just in case*

$$\begin{aligned} \mathcal{S} \;=\; \{r \in \mathcal{R} :\; & r \in Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S}), \\ & r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S}), \\ & r \notin Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S}), \\ & r \notin Excluded_{\mathcal{W},\mathcal{R}}(\mathcal{S})\}. \end{aligned}$$

With this simple change we're done. There's no need to change the definition of consequence. The following observation shows that the reasoner that can reason with the predicate *Reasonable* is a conservative generalization of the original one—the proof of the observation is in the Appendix:

**Observation 6** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context in which no Reasonable-formulas occur. Then there exists a context $c' = \langle \mathcal{W}, \mathcal{R} \rangle$ where Reasonable-formulas do occur such that*

*$X$ follows from $c$ if and only if $X$ follows from $c'$,*

*for all $X$ in which the predicate Reasonable doesn't occur.*

Our final rendering of Disagreement on Drugs is the context $c_{10} = \langle \mathcal{W}, \mathcal{R} \rangle$. The set $\mathcal{W}$ contains all the formulas the hard information of $c_9$ did, together with the following formulas, specifying the rules that the reasoner is committed to from the

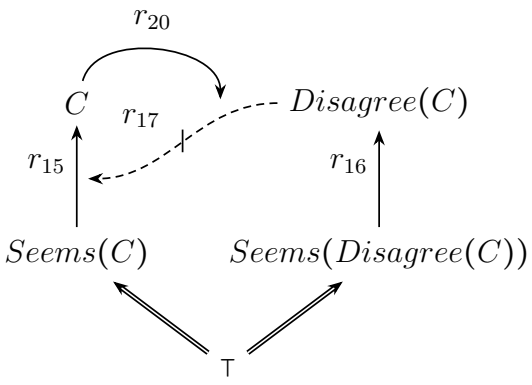



Figure 4.9: Disagreement with Evelyn

outset: $Reasonable(\mathfrak{r}_{12})$, $Reasonable(\mathfrak{r}_{13})$, $Reasonable(\mathfrak{r}_{15})$, $Reasonable(\mathfrak{r}_{18})$, and $Reasonable(\mathfrak{r}_{18})$. What conclusions does the reasoner draw from $c_{10}$? Well, there's only one proper scenario based on it, namely $\mathcal{S}_5 = \{r_{12}, r_{13}, r_{18}\}$, and so we have $c_{10} \mathrel{|\!\sim} L$ and $c_{10} \not\mathrel{|\!\sim} C$. Thus, our analysis suggests that the correct response to the scenario is to back off from the belief in conciliationism and be steadfast with regard to the belief in the existence of libertarian free will.

### 4.2.2 The (formal) inconsistency problem

Now let's refocus on the other half of the story in Double Disagreement. Consider the abridged version in which I reason my way to the conclusion that conciliationism is correct and then discover that my friend epistemologist Evelyn sincerely believes that it isn't. Our preliminary formalization made use of the formulas $Seems(C)$ and $Seems(Disagree(C))$, as well as the rules $r_{15} = \dfrac{Seems(C)}{C}$, $r_{16} = \dfrac{Seems(Disagree(C))}{Disagree(C)}$, and $r_{17} = \dfrac{Disagree(C)}{Out(\mathfrak{r}_{15})}$. (See Figure 4.6.) We found it lacking as it did not connect $C$ and $r_{17}$. Now, however, we know how to

connect them. Crucially, we must introduce a new rule $r_{20} = \dfrac{C}{Reasonable(\mathfrak{r}_{17})}$ which lets the reasoner conclude, by default, that $r_{17}$ is a reasonable rule to follow, once it concludes that $C$. And, then, we must also supplement the hard information with statements saying that $r_{15}$, $r_{16}$, and $r_{20}$ are reasonable. The result is the extended context $c_{11} = \langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W} = \{Seems(C), Seems(Disagree(C)), Reasonable(\mathfrak{r}_{15}), Reasonable(\mathfrak{r}_{16}), Reasonable(\mathfrak{r}_{20})\}$ and $\mathcal{R} = \{r_{15}, r_{16}, r_{17}, r_{23}\}$. It is depicted in Figure 4.9.

It turns out that there are no proper scenarios based on $c_{11}$.[26] This is bad news for the advocates of conciliatory views, since no proper scenarios means that we get $c_{11} \mathrel{|\!\sim} X$ for any formula $X$ whatsoever. Recall that, on our definition of consequence, a formula $X$ is said to follow from a context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ if and only if $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$ for every proper scenario based on $c$. When there are no proper scenarios, every formula satisfies the right-hand side of the biconditional vacuously, and so every formula follows from the context.[27]

[26] This can be verified by enumeration, going through all of the subsets of $\mathcal{R}$ one by one. Or we can save ourselves from the tedious exercise, restricting attention to the antecedently viable scenarios. Three things should be uncontroversial. First, given that $r_{16}$ is triggered by $\mathcal{W}$ and is not threatened by anything, it'd have to be included in *every* proper scenario based on $c_{11}$. Second, $r_{17}$ can be in a scenario only in case $r_{20}$ is—otherwise $r_{17}$ wouldn't be triggered. And third, $r_{20}$ can be in a proper scenario only in case $r_{15}$ is. There are four scenarios satisfying these three conditions: $\{r_{16}\}$, $\{r_{15}, r_{16}\}$, $\{r_{15}, r_{16}, r_{20}\}$, and $\{r_{15}, r_{16}, r_{17}, r_{20}\}$. The first three aren't proper as they fail to include all triggered default rules—e.g., $r_{15}$ is triggered in the context of $\{r_{16}\}$, but not included in it. And $\{r_{15}, r_{16}, r_{17}, r_{20}\}$ doesn't qualify as proper because it excludes one of its own elements, $r_{15}$.

[27] Notice that this outcome depends on the way we defined logical consequence, and, more specifically, that our definition requires that $X$ follows from *every* proper scenario based on the context. There's another natural and widely used definition of consequence—often called *credulous*—which requires that $X$ follows from *some* proper scenario based on the context, and one might wonder whether this alternative definition runs into this problem too. The answer is that it runs into a similar problem: When there are no proper scenarios based on a context, no formula whatsoever will follow from it on the credulous consequence. At first sight this might look like an improvement. Perhaps, it is correct to suspend judgment in response to the disagreement with Evelyn. However, it's important to notice that the reasoner's suspension of judgment is *universal*: For starters, it doesn't draw the conclusion $Disagree(C)$, or that there's genuine disagreement about conciliation-

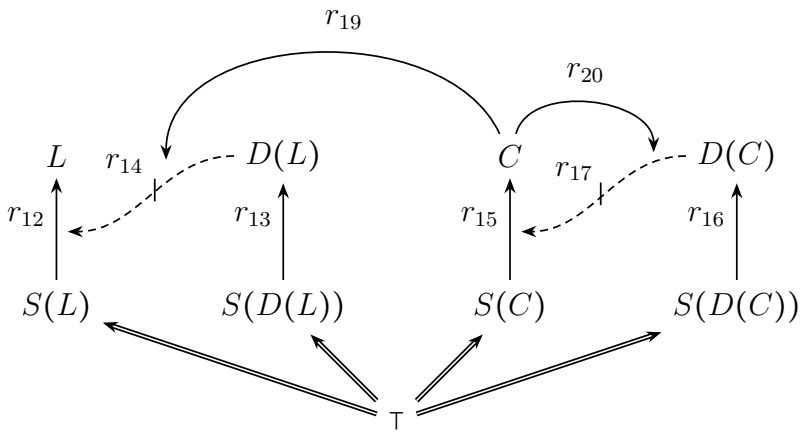



Figure 4.10: Double Disagreement, final

What's more, the situation doesn't get any better when we look at the context that captures the entire story recounted in Double Disagreement. Merging $c_{10}$ and $c_{11}$, we acquire our final formalization of the scenario, the context $c_{12}$—see Figure 4.10 for a graph. Yet again, there are no proper scenarios based on $c_{12}$, and we get $c_{12} \mathrel{\sim\!\!\!\vdash} X$ for all $X$. What this means is that our carefully designed model reasoner suggests that the correct conciliatory response to Double Disagreement is to conclude everything.

Recall that we started with an informal description of the problem of self-defeat for strong conciliatory views and noted that it seemed to split into two. The first problem appeared to be that conciliatory views issue inconsistent directives in possible scenarios of a certain structure; and the second that nobody could rationally hold a conciliatory view in the actual world. The problem that $c_{12}$ gives rise to seems to be naturally thought of as a formal version of the first problem: Advancing

ism, which we intuitively want it to draw in the case at hand. What's more, were we to supplement $c_{11}$ with information utterly unrelated to conciliationism and disagreement, such as $Perceive(R)$ with the corresponding rule $\dfrac{Perceive(R)}{R}$, the reasoner would suspended judgment on it too. So the alternative definition of consequence seems to fare no better than the one we are using.

a model (conciliatory) reasoner which suggests that concluding everything is the correct response to scenarios of a certain shape is much like advancing a (conciliatory) view that sometimes issues inconsistent directives. Now that we have this problem in front of us in plain sight, we don't only see that the worries about the behavior of conciliatory views in contexts involving disagreement about the correct way to disagree are legitimate, but can also start addressing them.

# Chapter 5: From default logic to formal argumentation

As we saw, our carefully designed model reasoner suggests that the correct conciliatory response to Double Disagreement—a scenario involving a disagreement about how to disagree—is to conclude everything. This might seem to corroborate Elga's (2010) conclusion that conciliationism is inherently flawed. However, conciliationism's turning on itself is only a part of the reason why the problematic result obtains. The other part is the way default logic handles what we'll call *self-defeating chains* or *vicious cycles* of rules. If we look at the context $c_{11}$, representing the second half of the Double Disagreement scenario—see Figure 5.1—it's very natural to single out the chain of rules $r_{15}$-$r_{20}$-$r_{17}$ as its problematic component. What happens is that the rule $r_{15}$ puts $r_{17}$ into place, via $r_{20}$, and thereby undermines its own support. This chain forms a vicious cycle, and it's just a general fact about default logic that it can't adequately handle contexts that contain cycles of this sort.[1]

Unfortunately, there's no straightforward fix to default logic that would let it generate meaningful consequences for $c_{11}$, $c_{12}$, and other contexts containing self-defeating chains. However, there's a roundabout way of getting to such consequences

[1]This aspect of default logic is discussed in (Horty 2012). Horty sees the existence of contexts containing vicious cycles as a technical problem—which it is—and hints at two general strategies for dealing with it. The first is to restrict the background language in a way that would preclude the possibility of default rules forming vicious cycles. And the second is to leave the language as is, but to modify the reasoner in a way that would let it generate meaningful conclusions in the presence of such cycles—see (Horty 2012, pp. 59–61). The first strategy is a no-go for us, since it would let us escape the unfortunate consequences of $c_{11}$ only by disallowing us to formalize the scenario as $c_{11}$. Our question is whether or not conciliationism's turning on itself results in inconsistent commands, and so adopting this strategy would be question-begging. But the second strategy is a different matter.

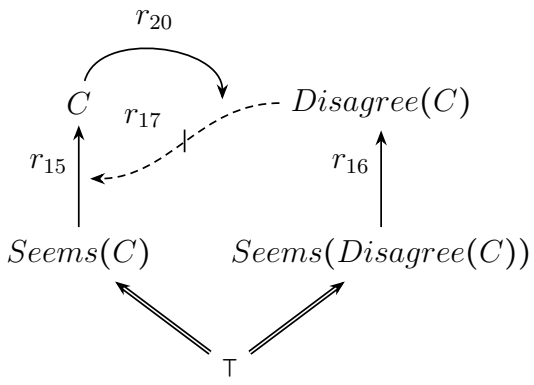



Figure 5.1: Disagreement with Evelyn, again

that takes a recourse through a different and more general formal framework called *abstract argumentation theory*. As the seminal work of Phan Minh Dung (1995) established, various logics for defeasible reasoning—including default logic—can be seen as special cases of argumentation theory. Part of this means is that one can formulate an *argumentation theory-based reasoner* that picks out the same consequence relation as the default logic-based reasoner we formulated above. What we're going to do next, then, is devise such a reasoner. Why? Well, because a simple tweak to it will give us a (more) sophisticated reasoner capable of handling theories with cycles in an adequate way.

The remainder of this chapter is structured as follows. The relatively technical Sections 5.1–5.3 introduce abstract argumentation, relate it to default logic, and devise the argumentation theory-based reasoner. Section 5.4 returns to Double Disagreement and explains how the sophisticated reasoner handles it.

## 5.1 Argument frameworks

While in default logic conclusions are derived on the bases of contexts, in argumentation theory they are derived on the basis of *argument* (or *argumentation*) *frameworks*. Formally, such frameworks are pairs of the from $\langle \mathcal{A}, \rightsquigarrow \rangle$, where the first element $\mathcal{A}$ is a set of arguments—really, a set whose elements can be anything—and the second element $\rightsquigarrow$ is a defeat relation among these arguments. Thus, for any two arguments $\mathcal{S}$ and $\mathcal{S}'$ in $\mathcal{A}$, the relation $\rightsquigarrow$ can tell us whether $\mathcal{S}$ defeats $\mathcal{S}'$ or not.[2] We'll denote argument frameworks with the letter $\mathcal{F}$. What argumentation theory, then, does is provide a number of sensible ways for selecting the *set of winning* (or *undefeated*) *arguments* of any given framework $\mathcal{F}$, the set which, then, determines the conclusions that can be drawn on the basis of $\mathcal{F}$.[3] In light of the fact that the frameworks we will focus on will be constructed from contexts, argumentation theory will let us determine the conclusions that can be drawn on the basis of any given context $c$.

Recall that our default logic-based reasoner relies on the notion of a proper scenario to determine the consequences of a context. This notion specifies something like the necessary and sufficient conditions for a default rule's counting as admissible or good—that the rule be *reasonable*, *triggered*, not *conflicted*, and not *excluded*—and the reasoner can be thought of as selecting the good rules in one single step. Notice that nothing would seem to stand in the way of selecting such

[2] Technically, the defeat relation $\rightsquigarrow$ is a subset of $\mathcal{A}\times\mathcal{A}$. Thus, argument frameworks are relatively simple mathematical objects, namely, directed graphs.

[3] I should note that, in general, a framework can have multiple winning argument sets.

rules in a more stepwise fashion. That is, instead of jumping from a context to the scenario containing all and only the admissible rules, we could first select *all* scenarios whose members satisfy the positive conditions—reasonable and triggered—and, then, later filter out the scenarios whose members do not satisfy the remaining negative conditions—conflicted and excluded. Or, to restate the idea using our formal notation, starting with a context $c = \langle \mathcal{W}, \mathcal{R} \rangle$, in the first step we would select all and only those scenarios $\mathcal{S} \subseteq \mathcal{R}$ such that, for every $r$ in $\mathcal{S}$,

$$\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Reasonable(r) \text{ and } \mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Premise[r],$$

and, in the second step, we would filter out all of those scenarios $\mathcal{S}$ for which it holds that there's some $r$ in $\mathcal{S}$ such that

$$\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r] \text{ or } \mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(\mathfrak{r}).$$

After taking these two steps, we'd have access to all and only the good default rules. Now, the application of argumentation to contexts can be naturally thought of as proceeding in these two steps. The scenarios selected in the first step will just be the arguments of the argument framework based on the given context. And the scenarios that remain standing after the second step will be the winning (or the undefeated) arguments of the framework. So here's our definition of an argument based on a context:

**Definition 5.1 (Arguments)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ a scenario based on it, $\mathcal{S} \subseteq \mathcal{R}$. Then $\mathcal{S}$ is an* argument *based on $c$ just in case $\mathcal{S} \subseteq Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S})$ and $\mathcal{S} \subseteq Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. The set of arguments based on $c$ is the set $Arguments(c)$ $= \{\mathcal{S} \subseteq \mathcal{R} : \mathcal{S}$ is an argument based on $c\}$.*

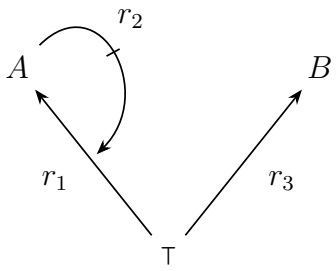



Figure 5.2: Sample context with a vicious cycle

To see this definition at work, let's apply it to an abstract toy example. Consider the context $c_{13} = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ contains the statements $\mathit{Reasonable}(\mathfrak{r}_1)$, $\mathit{Reasonable}(\mathfrak{r}_2)$, and $\mathit{Reasonable}(\mathfrak{r}_3)$ and $\mathcal{R}$ contains the rules $r_1 = \frac{\top}{A}$, $r_2 = \frac{A}{\mathit{Out}(\mathfrak{r}_1)}$, and $r_3 = \frac{\top}{B}$. The context is depicted in Figure 5.2. As one can easily verify, there are eight scenarios based on $c_{13}$, namely,

$$\mathcal{S}_0 = \varnothing, \qquad \mathcal{S}_4 = \{r_1, r_2\},$$
$$\mathcal{S}_1 = \{r_1\}, \quad \mathcal{S}_5 = \{r_1, r_3\},$$
$$\mathcal{S}_2 = \{r_2\}, \quad \mathcal{S}_6 = \{r_2, r_3\},$$
$$\mathcal{S}_3 = \{r_3\}, \quad \mathcal{S}_7 = \{r_1, r_2, r_3\}.$$

Two of these scenarios, $\mathcal{S}_2$ and $\mathcal{S}_6$, fail to qualify as arguments. And the reason is that they both contain an element, the rule $r_2$, that's not triggered. Indeed, one glance at the graph depicting $c_{13}$ in Figure 5.2 is enough to see that $r_2$ won't be triggered in any scenario based on $c_{13}$ that fails to contain $r_1$. This leaves us with six arguments that will comprise the first element $\mathcal{A}$ of the argument framework based on $c$.

The next step is to specify the conditions under which one argument defeats

another, and that's just what our next definition does.

**Definition 5.2 (Defeat)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ and $\mathcal{S}'$ two arguments based on it. Then $\mathcal{S}$* defeats *$\mathcal{S}'$, written as $\mathcal{S} \rightsquigarrow \mathcal{S}'$, if and only if there is some rule $r \in \mathcal{S}'$ such that either $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$, or $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(\mathfrak{r})$.*

Notice how the core ideas behind the notions of conflicted and excluded rules are repurposed in this definition: A rule $r$ came out conflicted in the context of a scenario $\mathcal{S}$ just in case $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$, and now an argument $\mathcal{S}$ defeats another argument $\mathcal{S}'$ if there is a rule $r$ in $\mathcal{S}'$ such that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$; and similarly for exclusion. Now let's see the definition at work when applied to the context $c_{13}$, starting with the arguments $\mathcal{S}_7 = \{ r_1 = \frac{\top}{A}, \ r_2 = \frac{A}{Out(\mathfrak{r}_1)}, \ r_3 = \frac{\top}{B} \}$ and $\mathcal{S}_1 = \{ r_1 = \frac{\top}{A} \}$. Notice that the set $Conclusion[\mathcal{S}_7]$ entails $Out(\mathfrak{r}_1)$ and that $r_1$ is in $\mathcal{S}_1$. So, according to the definition, $\mathcal{S}_7$ defeats $\mathcal{S}_1$. What's more, $r_1$ is an element of $\mathcal{S}_7$, meaning that the argument $\mathcal{S}_7$ self-defeats.[4]

Now we have all that's needed to specify how to construct argument frameworks from contexts.

**Definition 5.3 (Context-based argument frameworks)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be an epistemic context. Then the* argument framework *$\mathcal{F}(c)$ based on $c$ is the pair $\langle \mathcal{A}, \rightsquigarrow \rangle$*

[4]It may be worth pointing out that any argument $\mathcal{S}$ based on some context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ that happens to contradict the hard information $\mathcal{W}$ will end up self-defeating. It's easy to see why this holds: By assumption, $\mathcal{W}$ together with $Conclusion[\mathcal{S}]$ is inconsistent, that is, $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \bot$. But given that inconsistent sets entail all formulas, it must hold that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$ for every $r$ in $\mathcal{S}$. And this is enough to conclude that $\mathcal{S} \rightsquigarrow \mathcal{S}$.

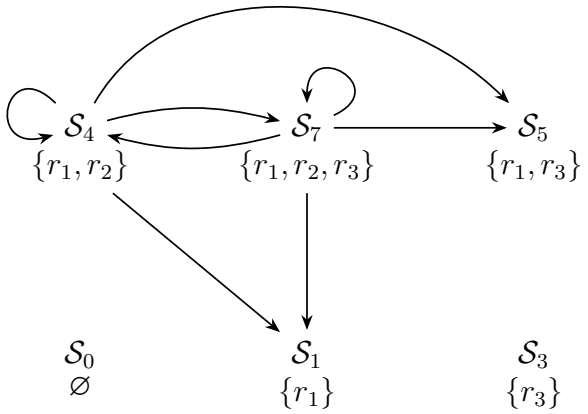



Figure 5.3: Argument framework for the sample context

*where* $\mathcal{A} = \mathit{Arguments}(c)$ *and* $\rightsquigarrow$ *is the set* $\{(\mathcal{S}, \mathcal{S}') \in \mathcal{A} \times \mathcal{A} : \mathcal{S}$ *defeats* $\mathcal{S}'\}$.

The graph representing the argument framework $\mathcal{F}(c_{13})$ constructed from $c_{13}$ is depicted in Figure 5.3. Here's how it should be read. The nodes of the graph, $\mathcal{S}_0$, $\mathcal{S}_1$, and so on, are the arguments in $\mathcal{A}$, and the pointed edges between them are the relations of defeat. Nodes with edges pointing to themselves are the arguments that happen to self-defeat.

It'll also be useful to introduce some shorthand notation here. Let $\mathcal{F} = \langle \mathcal{A}, \rightsquigarrow \rangle$ be an arbitrary argument framework and $\Gamma$ and $\Gamma'$ two sets of arguments from $\mathcal{A}$. When there's an argument $\mathcal{S}$ in $\Gamma$ that defeats some argument $\mathcal{S}'$ from $\mathcal{A}$, we will write $\Gamma \rightsquigarrow \mathcal{S}'$; and when there's a pair of arguments $\mathcal{S}$ and $\mathcal{S}'$ such that $\mathcal{S}$ is in $\Gamma$, $\mathcal{S}'$ is in $\Gamma'$, and $\mathcal{S} \rightsquigarrow \mathcal{S}'$, we will write $\Gamma \rightsquigarrow \Gamma'$. As an illustration, in the case of $\mathcal{F}(c_{13})$, we have $\{\mathcal{S}_0, \mathcal{S}_7\} \rightsquigarrow \mathcal{S}_5$, while we do not have $\{\mathcal{S}_0, \mathcal{S}_7\} \rightsquigarrow \mathcal{S}_3$; and we have $\{\mathcal{S}_0, \mathcal{S}_7\} \rightsquigarrow \{\mathcal{S}_1, \mathcal{S}_3, \mathcal{S}_4\}$, while we do not have $\{\mathcal{S}_0, \mathcal{S}_7\} \rightsquigarrow \{\mathcal{S}_3\}$. Further, when there's no argument $\mathcal{S}$ in $\Gamma$ such that $\mathcal{S}$ defeats $\mathcal{S}'$, we will write $\Gamma \not\rightsquigarrow \mathcal{S}'$; and when there's no pair of arguments $\mathcal{S}$ and $\mathcal{S}'$ such that $\mathcal{S}$ is in $\Gamma$, $\mathcal{S}'$ is in $\Gamma'$, and $\mathcal{S} \rightsquigarrow \mathcal{S}'$, we will

write $\Gamma \nsim \Gamma'$.

## 5.2 Selecting winning arguments

In this section, we turn to argumentation theory proper. As mentioned above, in it the winning argument set of a framework—or, rather, sets, as one framework can have multiple winning sets—are selected *only* on the basis of the defeat relation among arguments. In the literature, the collection of definitions that lets one select winning sets is standardly referred to as *admissibility semantics*. You may find it helpful to think of this semantics as serving a function that's similar to the one served by the notion of a proper scenario in the context of default logic. There's one important difference, however. Where default logic didn't offer any choice, the admissibility semantics provides a number of different sensible ways of selecting winning arguments. We will focus on two such here, beginning with what we'll call *stability semantics*.

**Definition 5.4 (Stability semantics)** *Let $\mathcal{F} = \langle \mathcal{A}, \rightsquigarrow \rangle$ be an argument framework and $\Gamma$ a set of arguments from $\mathcal{A}$. Then:*

*(i)* $\Gamma$ *is* conflict-free *if and only if there are no two arguments $\mathcal{S}$, $\mathcal{S}'$ in $\Gamma$ such that $\mathcal{S} \rightsquigarrow \mathcal{S}'$,*

*(ii)* $\Gamma$ *is* stable*, or a* stable extension of $\mathcal{F}$*, if and only if*

*(1)* $\Gamma$ *is conflict-free, and*

*(2)* $\Gamma$ *defeats all the arguments that are not in it, that is, for all* $\mathcal{S} \in \mathcal{A}\backslash\Gamma$*,* $\Gamma \rightsquigarrow \mathcal{S}$*.*

Stability semantics is closely related to default logic, and we will state the precise connection between the two in a moment. For now, let me just point out that there are no stable argument sets based on the framework $\mathcal{F}(c_{13})$, just like there are no proper scenarios based on the context $c_{13}$. As you may have guessed, this is due to the self-defeating chain of rules $r_1$-$r_2$.

The alternative to the stability semantics we are going to make use of is what we'll call *preference semantics*:

**Definition 5.5 (Preference semantics)** *Let* $\mathcal{F} = \langle \mathcal{A}, \rightsquigarrow \rangle$ *be an argument framework and* $\Gamma$ *a set of arguments from* $\mathcal{A}$*. Then:*

*(i)* $\Gamma$ *is* conflict-free *if and only if there are no arguments* $\mathcal{S}$*,* $\mathcal{S}'$ *in* $\Gamma$ *such that* $\mathcal{S} \rightsquigarrow \mathcal{S}'$*.*

*(ii) An argument* $\mathcal{S}$ *in* $\mathcal{A}$ *is* defended *by* $\Gamma$ *if and only if, for all* $\mathcal{S}'$ *(with* $\mathcal{S}'$ *in* $\mathcal{A}\backslash\Gamma$*) such that* $\mathcal{S}' \rightsquigarrow \mathcal{S}$*, we have* $\Gamma \rightsquigarrow \mathcal{S}'$*.*

*(iii)* $\Gamma$ *is* a complete extension *of* $\mathcal{F}$ *if and only if*

*(1)* $\Gamma$ *is conflict-free, and*

*(2)* $\Gamma$ *contains all of the arguments it defends.*

*(iv)* $\Gamma$ *is* preferred*, or a* preferred extension of $\mathcal{F}$*, if and only if* $\Gamma$ *is a* maximal *complete extension of* $\mathcal{F}$*, that is, if and only if there's no other complete*

*extension* $\Gamma'$ *of* $\mathcal{F}$ *such that* $\Gamma \subset \Gamma'$.

Now let's apply this definition to $\mathcal{F}(c_{13})$. The largest conflict-free set of arguments is $\{\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_3, \mathcal{S}_5\}$. However, it does not defend all of its members: Both $\mathcal{S}_4$ and $\mathcal{S}_7$ defeat $\mathcal{S}_5$, and there's nothing in the set that defeats either of them. There are a few sets that are both conflict-free and defend all of their members, namely, $\varnothing$, $\{\mathcal{S}_0\}$, $\{\mathcal{S}_3\}$, and $\{\mathcal{S}_0, \mathcal{S}_3\}$. However, only one of them qualifies as a complete extension, namely, $\{\mathcal{S}_0, \mathcal{S}_3\}$. Why do the other sets fail to qualify? Take $\{\mathcal{S}_0\}$ as an example. A complete set is supposed to contain all the arguments it defends, and a minute's reflection should suffice to see that $\{\mathcal{S}_0\}$ does not contain an argument it defends, namely, $\mathcal{S}_3$. What's more, the set $\{\mathcal{S}_0, \mathcal{S}_3\}$ also happens to be the unique preferred extension of $\mathcal{F}(c_{13})$.

Abstract as it might look, preference semantics has a clear intuitive rationale. Any conflict-free set of arguments that defends itself—that is, any complete extension—is a desirable state to occupy: It is consistent, and it has a rejoinder to every attack on it. And there's a clear sense in which a preferred set of arguments is an even more desirable state to occupy: It is still consistent; it still has a rejoinder to every attack on it; and it's also as big of an argument set of this sort as there can be. So if we're looking to select a winning argument set of some framework $\mathcal{F}$, then preferred extensions seem to be very natural candidates.[5]

[5]It's worth pointing out that argumentation theory offers alternative ways of selecting winning sets of arguments too. The one that's at least as important as the stability and preference semantics results in selecting the so-called *grounded extensions*. A grounded extension is dual to preferred extensions: It is the (set theoretically) minimal complete extensions of a given argument framework. Just as it is in the case of preferred extensions, and not in the case of the stable ones, grounded extensions are guaranteed to exist, whether or not the underlying argument frameworks contains self-defeating arguments. Given that our goal here is to formulate a defeasible reasoner that draws

We have given labels to various kinds of sets of arguments, but we still haven't made clear the conditions under which a formula follows from such a set. The following definition rectifies the omission:

**Definition 5.6** *Where $\mathcal{F}(c)$ is an argument framework constructed from some context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ and $\Gamma$ is a set of arguments based on $\mathcal{F}(c)$, a statement $X$ is a conclusion of $\Gamma$ if and only if there is some argument $\mathcal{S}$ in $\Gamma$ such that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$.*

With this definition in place, we have all the elements we need to define two distinct consequence relations. Both will specify when a formula $X$ follows from a context $c$. Both take a circuitous route, utilizing the resources of argumentation theory. The only difference between them is that the first relies on stability semantics, while the second relies on preference semantics:

**Definition 5.7 (Consequence, stable)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context. Then the statement $X$ follows from $c$ according to stability semantics—written $c \mathrel{|\!\sim}_s X$—just in case it is a conclusion of every* stable extension *of the argument framework $\mathcal{F}(c)$.*

**Definition 5.8 (Consequence, preferred)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context. Then the statement $X$ follows from $c$ according to preference semantics—written $c \mathrel{|\!\sim}_p X$—just in case it is a conclusion of every* preferred extension *of the argument framework $\mathcal{F}(c)$.*

sensible conclusions in the presence of self-defeating chains of rules, nothing of importance would change had we opted for the grounded extensions instead of the preferred ones. It's noteworthy that the grounded extension and the unique preferred extension of $\mathcal{F}(c_{13})$ coincide. There are argument frameworks whose grounded and preferred extensions come apart, but, given what our aims are in this chapter, we need not concern ourselves with them.

The promised connection between default logic and stability semantics can now be stated in the form of an observation—the proof of which can be found in the Appendix:

**Observation 7** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $X$ an arbitrary formula. Then $X$ follows from $c$ in default logic, $c \mathrel{|\!\sim} X$, if and only if $X$ follows from $c$ according to stability semantics, $c \mathrel{|\!\sim}_s X$.*

Thus, stability semantics expresses default logic in argumentation-theoretic terms, and so inherits both its virtues and its vices. We have seen one of its vices: Default logic collapses in the presence of self-defeating chains. Having the tools of argumentation theory at our disposal, we can get an insight into why this happens. When an argument contains a self-defeating chain of rules, it ends up self-defeating. Recall that stable extensions are conflict-free argument sets that defeat all the arguments that aren't in them. Now suppose that we have some argument framework $\mathcal{F} = \langle \mathcal{A}, \rightsquigarrow \rangle$, that $\mathcal{S}$ is some self-defeating argument from $\mathcal{A}$, and that $\Gamma$ is a would-be stable extension of $\mathcal{F}$. How would $\Gamma$ relate to $\mathcal{S}$? Clearly, $\Gamma$ can't include $\mathcal{S}$. Otherwise, it wouldn't be conflict-free. So $\Gamma$ must defeat $\mathcal{S}$. However, unless $\mathcal{F}$ is based on a context of a very particular structure, there's just not going to be an *independent* argument in $\Gamma$—that is, an argument that's neither a subset, nor a superset of $\mathcal{S}$—that would defeat $\mathcal{S}$. And if there's no such argument, $\Gamma$ can't qualify as a stable extension. The rather demanding character of stability semantics becomes manifest once we contrast it with preference semantics. Where preference semantics only requires that a winning argument set has a rejoinder to every attack on it,

stability semantics requires that a winning set attacks *every* argument that isn't in it.

I propose that we switch from the consequence relation picked out by both default logics and the stability semantics to the consequence relation picked out by the preference semantics. The move is not ad hoc, because there is a clear sense in which preference semantics is a conservative generalization of stability semantics, and, thus, also of default logic. This sense is captured by the following observation—the proof of which can, again, be found in the Appendix:

**Observation 8** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{F}(c) = \langle \mathcal{A}, \rightsquigarrow \rangle$ an argument framework constructed from it. If $\mathcal{F}(c)$ does not contain either odd cycles of defeat or infinite chains of defeat, then $c \mathrel{\sim\!\!\!\!\vdash}_s X$ if and only if $c \mathrel{\sim\!\!\!\!\vdash}_p X$.*

One consequence of this is that preference semantics gives the same results as default logic in all contexts that do not contain any self-defeating chains of rules.[6] When such chains are present, default logic will return the trivial set, while preference semantics will return more reasonable consequences. The toy context $c_{13}$ is a case in point. When we run default logic on it, we get $c_{13} \mathrel{\sim\!\!\!\!\vdash} X$ for any formula $X$ whatsoever, while, when we rely on the preference semantics, we get the more reasonable $c_{13} \mathrel{\sim\!\!\!\!\vdash}_p B$ and $c_{13} \not\mathrel{\sim\!\!\!\!\vdash}_p A$. Here, as well as in general, preference semantics effectively disregards the self-defeating chains of rules and draws conclusions on the basis of only those rules that are independent of such chains.

[6] How exactly does this follow from the observation? Well, the presence of a self-defeating chain of rules in $c$ typically means that the argument framework $\mathcal{F}(c)$ constructed from $c$ contains at least one self-defeating argument—this doesn't happen only in those cases where not a single arguments of $\mathcal{F}(c)$ subsumes the chain. But self-defeating argument are (odd) cycles of defeat. So if $\mathcal{F}(c)$ has no odd cycles of defeat, then $c$ cannot contain self-defeating chains of rules.

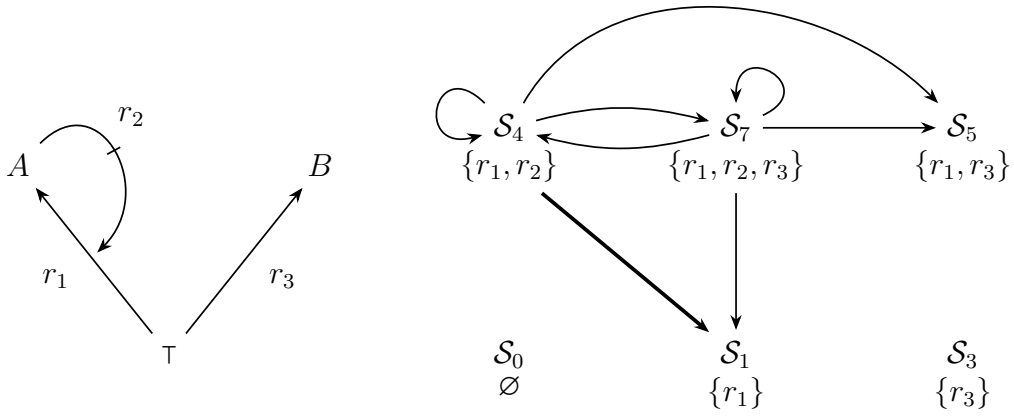



Figure 5.4: Context $c_{13}$ and the corresponding framework $\mathcal{F}(c_{13})$

At this point we could return to Double Disagreement and the problem it gave rise to. But before we do that, I want to introduce an alternative way of thinking about defeat between arguments.

## 5.3 Minimal arguments and basic defeat

There's an intuitive sense in which some arguments are basic and others are not. To see this, let's take another look at our running example $c_{13}$ and the corresponding argumentation framework $\mathcal{F}(c_{13})$, depicted side by side in Figure 5.4 for convenience. The scenarios $\mathcal{S}_1 = \{r_1\}$, $\mathcal{S}_3 = \{r_3\}$, $\mathcal{S}_4 = \{r_1, r_2\}$, $\mathcal{S}_5 = \{r_1, r_3\}$, and $\mathcal{S}_7 = \{r_1, r_2, r_3\}$ all qualify as arguments based on $c_{13}$, but there seems to be a qualitative difference between the first three, $\mathcal{S}_1$, $\mathcal{S}_3$, and $\mathcal{S}_4$, on the one hand, and $\mathcal{S}_5$ and $\mathcal{S}_7$, on the other. For one, $\mathcal{S}_1$, $\mathcal{S}_3$, and $\mathcal{S}_4$ are the (set-theoretically) smallest arguments allowing us to derive, respectively, $A$, $Out(\mathfrak{r}_1)$, and $B$. Let's compare $\mathcal{S}_1$ and $\mathcal{S}_7$. While we have both $\mathcal{W} \cup Conclusion[\mathcal{S}_1] \vdash A$ and $\mathcal{W} \cup Conclusion[\mathcal{S}_7] \vdash A$,

only in the case of $\mathcal{S}_1$ can we say that there's no smaller argument that would let us derive $A$. Also, $\mathcal{S}_5$ and $\mathcal{S}_7$ can be naturally thought of as aggregates of the basic arguments: $\mathcal{S}_5$ combines $\mathcal{S}_1$ and $\mathcal{S}_3$, while $\mathcal{S}_7$ combines $\mathcal{S}_1$, $\mathcal{S}_3$, and $\mathcal{S}_4$.

What's more, we can identify a correspondingly basic defeat relation between arguments. There is, again, a clear intuitive sense in which the real action happens between the smallest argument to the conclusion $Out(\mathfrak{r}_1)$, namely, $\mathcal{S}_4$, and the smallest argument containing $r_1$, namely, $\mathcal{S}_1$. In the graph depicting the framework $\mathcal{F}(c_{13})$ in Figure 5.4, this relation is represented by the highlighted arrow. The defeat relations obtaining between the other arguments seem to depend on this one, and in a way we can capture precisely: For any two arguments $\mathcal{S}$, $\mathcal{S}'$ in $\mathcal{F}(c_{13})$, we have $\mathcal{S} \rightsquigarrow \mathcal{S}'$ only if the basic argument $\mathcal{S}_4$ is a part of $\mathcal{S}$, $\mathcal{S}_4 \subseteq \mathcal{S}$, and the basic argument $\mathcal{S}_1$ is a part of $\mathcal{S}'$, $\mathcal{S}_1 \sqsubseteq \mathcal{S}'$.

The goal of this section is to capture this intuitive sense of basicness—basic arguments and basic defeat—in a mathematically precise way, and to show how it leads to an alternative (yet extensionally equivalent) characterization of the defeat relation between arguments.

We'll be exploring an alternative way of constructing argument frameworks from contexts. Nothing changes with regard to arguments. In specifying the relation of defeat, the first step is to find a way to select those arguments from $c$ that appear basic intuitively. Notice that an argument that's basic with respect to one rule or one formula doesn't have to count as basic with respect to another default rule or formula. Consider our running example again. The scenario $\mathcal{S}_4 = \{r_1, r_2\}$ seems basic with respect to both the rule $r_2$ and the formula $Out(\mathfrak{r}_1)$, since there's no smaller

argument that would either contain $r_2$, or let us derive $Out(\mathfrak{r}_1)$. However, $\mathcal{S}_4$ is not basic with respect to $r_1$ and the formula $A$, since there's the smaller $\mathcal{S}_1 = \{r_1\}$ which contains $r_1$ and let's us derive $A$. What this means is that the formal notion capturing the intuitive idea of basicness must be relativized, to a rule or a formula. But otherwise, the basic, or *minimal*, arguments just are the (set-theoretically) smallest arguments we can find:

**Definition 5.9 (Minimal arguments, with respect to rules)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $r$ a rule from $\mathcal{R}$. Then the* $r$-minimal arguments*, in the context of $c$, are those arguments that belong to the set*

$$\begin{aligned} Minimal_{\mathcal{F}(c)}(r) \;=\; \{\mathcal{S} \in Arguments(c) :\; & r \in \mathcal{S} \text{ and} \\ & \nexists\, \mathcal{S}' \in Arguments(c) \text{ such that} \\ & \quad (1)\ r \in \mathcal{S}' \text{ and} \\ & \quad (2)\ \mathcal{S}' \subset \mathcal{S}\}. \end{aligned}$$

**Definition 5.10 (Minimal arguments, with respect to formulas)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $X$ a formula of our language. Then the* $X$-minimal arguments*, in the context of $c$, are those arguments that belong to the set*

$$\begin{aligned} Minimal_{\mathcal{F}(c)}(X) = \{\mathcal{S} \in Arguments(c) :\; & \mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X \text{ and} \\ & \nexists\, \mathcal{S}' \in Arguments(c) \text{ such that} \\ & \quad (1)\ \mathcal{W} \cup Conclusion[\mathcal{S}'] \vdash X \text{ and} \\ & \quad (2)\ \mathcal{S}' \subset \mathcal{S}\}. \end{aligned}$$

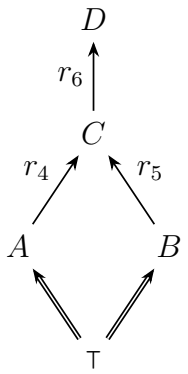



Figure 5.5: Multiple basic arguments

The plural in the definitions is not accidental. In general, there can be multiple $r$- or $X$-minimal arguments, as our next (abstract) example makes plain. Consider the context $c_{14} = \langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W} = \{A,\ B,\ \mathit{Reasonable}(\mathfrak{r}_4),\ \mathit{Reasonable}(\mathfrak{r}_5),\ \mathit{Reasonable}(\mathfrak{r}_6)\}$ and $\mathcal{R}$ consisting of the rules $r_4 = \frac{A}{C}$, $r_5 = \frac{B}{C}$, and $r_6 = \frac{C}{D}$. One glance at the inference graph depicting this context—see Figure 5.5—is enough to realize that there are two alternative ways of reaching $D$, by means of the chain $r_5$-$r_6$ and by means of the chain $r_4$-$r_6$. And, indeed, when we apply Definition 5.10 to this context, two arguments come out as $D$-minimal, $\{r_4, r_6\}$ and $\{r_5, r_6\}$. What's more, the same two arguments qualify as $r_6$-minimal.[7]

Now let's turn to basic defeat. While our next definition might look somewhat involved, all it does is capture the intuition we started with. For one argument $\mathcal{S}$ to *basic-defeat* another argument $\mathcal{S}'$, there has to be a rule $r$ such that $\mathcal{S}'$ is $r$-minimal and $\mathcal{S}$ is either $\neg \mathit{Conclusion}[r]$- or $\mathit{Out}(\mathfrak{r})$-minimal.

[7]In all of the examples we considered thus far, the $r$-minimal arguments coincided with the $\mathit{Conclusion}[r]$-minimal ones. To see that this doesn't hold in general, consider some context $c$ containing two default rules with the same conclusion, say, $r_1 = \frac{\top}{A}$ and $r_1' = \frac{\top}{A}$. Here will *not* have $\mathit{Minimal}_{\mathcal{F}(c)}(A) \neq \mathit{Mininimal}_{\mathcal{F}(c)}(r_1) \neq \mathit{Mininimal}_{\mathcal{F}(c)}(r_1')$, unless $\mathcal{W} \vdash A$.

**Definition 5.11 (Basic defeat)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be an ordinary context and $\mathcal{S}$ and $\mathcal{S}'$ two arguments of the framework $\mathcal{F}(c)$. Then $\mathcal{S}$ basic-defeats $\mathcal{S}'$, written as $\mathcal{S} \rightsquigarrow_b \mathcal{S}'$, if and only if there is some rule $r \in \mathcal{R}$ such that*

*(i) $\mathcal{S}'$ is in $Minimal_{\mathcal{F}(c)}(r)$ and*

*(ii) either (1) or (2):*

*(1) $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$ and $\mathcal{S}$ is in $Minimal_{\mathcal{F}(c)}(\neg Conclusion[r])$,*

*(2) $\mathcal{W} \cup Conclsuion[\mathcal{S}] \vdash Out(\mathfrak{r})$ and $S$ is in $Minimal_{\mathcal{F}(c)}(Out(\mathfrak{r}))$.*

Returning to the running example $c_{13}$, it's easy to see that the only two arguments that stand in the basic-defeat relation are $\mathcal{S}_4$ and $\mathcal{S}_1$. For $\mathcal{S}_4$ is the only element of the set $Minimal_{\mathcal{F}(c_{13})}(Out(\mathfrak{r}_1))$ and $\mathcal{S}_1$ is the only element of the set $Minimal_{\mathcal{F}(c_{13})}(r_1)$. In light of the fact that $Conclusion[\mathcal{S}_4]$ entails $Out(\mathfrak{r}_1)$, we have $\mathcal{S}_4 \rightsquigarrow_b \mathcal{S}_1$.

With the basic defeat relation in place, we can extrapolate it to arguments of arbitrary complexity, as follows:

**Definition 5.12 (Defeat, alternative definition)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ and $\mathcal{S}'$ two arguments based on it. Then $\mathcal{S}$ defeats $\mathcal{S}'$, written $\mathcal{S} \rightsquigarrow_a \mathcal{S}'$, if and only if there is an $\mathcal{S}'' \subseteq \mathcal{S}$ and an $\mathcal{S}''' \subseteq \mathcal{S}'$ such that $\mathcal{S}'' \rightsquigarrow_b \mathcal{S}'''$.*

It's not difficult to see that this definition lets us reestablish the defeat relations of the argument framework $\mathcal{F}(c_{13})$. This is not a coincidence, as our next observation makes clear—the proof is in the Appendix:

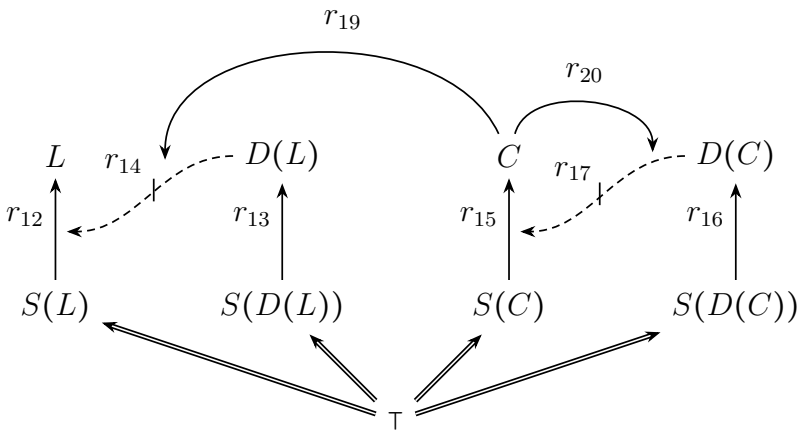



Figure 5.6: Double Disagreement, again

**Observation 9** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ and $\mathcal{S}'$ arguments from the argument framework $\mathcal{F}(c)$ constructed from it. Then $\mathcal{S}$ defeats $\mathcal{S}'$, according to Definition 5.2, $\mathcal{S} \rightsquigarrow \mathcal{S}'$, if and only if $\mathcal{S}$ defeats $\mathcal{S}'$, according to Definition 5.12, $\mathcal{S} \rightsquigarrow_a \mathcal{S}'$.*

Since Definitions 5.2 and 5.12 characterize the same notion, we can go back and forth between the two ways of thinking about defeat. That being said, in what follows we'll mostly work with the alternative characterization, as it is the easier to get an intuitive grip on.[8]

## 5.4 Back to Double Disagreement

Recall the Double Disagreement scenario in which the reasoning agent is confronted with two disagreeing peers: The metaphysician Milo disagrees with her about the existence of free will, and the epistemologist Evelyn disagrees with her about the truth of conciliationism. We encoded this scenario in the context $c_{12}$, depicted

[8] Also, in Chapter 6, we'll be adding another twist to our model, letting one rule $r$ support its conclusion to greater degree than another rule $r'$ supports its conclusion, and there we will be relying on the alternative definition.

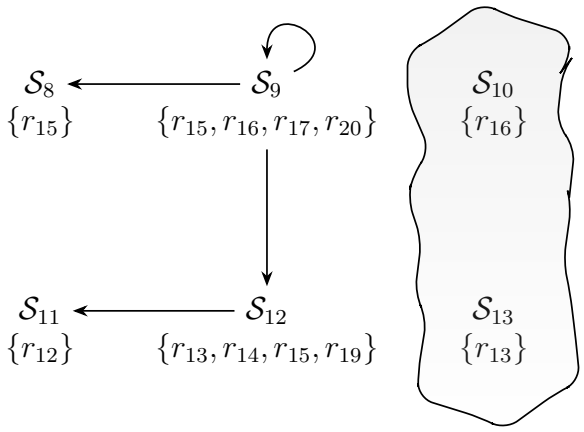



Figure 5.7: Core arguments from $\mathcal{F}(c_{12})$

again in Figure 5.6. Our original model reasoner let us derive any formula $X$ on the basis of this $c_{12}$, suggesting that the correct response to the scenario is to conclude everything. We then observed that this unfortunate result obtains due to the presence of the vicious cycles of rules $r_{15}$-$r_{20}$-$r_{17}$ paired with the fact that default logic isn't well-suited for dealing with cycles of this sort.

Now let's see if the more sophisticated reasoner—our conservative generalization of default logic—does any better. Since $c_{12}$ is a fairly complex context, it would be difficult to explore the entire argumentation framework $\mathcal{F}(c_{12})$ constructed from it. Luckily, however, we do not need to do that. For our purposes, it will suffice to look at its fragment. More specifically, we'll be interested in the fragment containing its most informative minimal arguments, together with the defeat relations between them. This includes:

the $C$-minimal argument $\mathcal{S}_8 = \{r_{15}\}$;

the $Out(\mathfrak{r}_{15})$-minimal $\mathcal{S}_9 = \{r_{15}, r_{16}, r_{17}, r_{20}\}$, which defeats $\mathcal{S}_8$, itself, and $\mathcal{S}_{12}$;

the $Disagree(C)$-minimal $\mathcal{S}_{10} = \{r_{16}\}$;

the $L$-minimal $\mathcal{S}_{11} = \{r_{12}\}$;

the $Out_{12}$-minimal $\mathcal{S}_{12} = \{r_{13}, r_{14}, r_{15}, r_{19}\}$, which defeats $\mathcal{S}_{11}$; and

the $Disagree(L)$-minimal argument $\mathcal{S}_{13} = \{r_{13}\}$.

Figure 5.7 depicts the fragment graphically. The lightly shaded region represents the preferred extension of $\mathcal{F}(c_{12})$—or, rather, its relevant fragment. It's not difficult to see why the defeat relations obtain: The argument $\mathcal{S}_9$ supports the formula $Out(\mathfrak{r}_{15})$, suggesting that the rule $r_{15}$ be taken out of consideration, and this rule is an element of $\mathcal{S}_8$, $\mathcal{S}_9$ itself, and $\mathcal{S}_{12}$. Similarly, the argument $\mathcal{S}_{12}$ supports the formula $Out(\mathfrak{r}_{12})$, suggesting that $r_{12}$ be taken out of consideration, and this rule is an element of $\mathcal{S}_{11}$. As the picture makes clear, the arguments $\mathcal{S}_8$ and $\mathcal{S}_{11}$ are not included in the preferred extension of $\mathcal{F}(c_{12})$. And this means that neither $C$, nor $L$ follow from the context, or that we get $c_{12} \not\sim_p C$ and $c_{12} \not\sim_p L$.

(It might also be worthwhile to discuss the manner in which the new reasoner handles vicious cycles in terms of rules, as opposed to in terms of arguments: What it does, in effect, is disregard all the rules involved in a vicious cycle ($r_{15}$, $r_{17}$, and $r_{20}$), as well as the rules that are not part of the cycle themselves, but are affected by rules that are, either directly or indirectly ($r_{19}$ and $r_{12}$). And for the rest, the new reasoner proceeds like the original one, drawing conclusions on the basis of the remaining rules that are reasonable, triggered, not conflicted, and not excluded. So the only rules that end up getting selected in the case of $c_{12}$ are $r_{13}$ and $r_{16}$.)

The fact that neither $C$, nor $L$ follow from $c_{12}$ means that the new reasoner—henceforth *the reasoner*, without qualification—suggests that the correct response to Double Disagreement is to abandon both the belief in conciliationism and the belief in the existence of free will. What should we make of this response? Well, the first thing to note is that it is perfectly consistent. The conciliatory view our model captures is rather extreme, and yet it doesn't issue inconsistent directives in situations involving disagreement about the correct way to disagreement. It can, thus, serve as an existence proof showing that Elga's (2010) conclusion regarding conciliatory views is mistaken. It's not the case that such views lead to inconsistency, when they turn on themselves.

But the fact that a view doesn't issue inconsistent directives, of course, doesn't make the view plausible, let alone show that it is correct. And there would seem to be two reasons to feel uneasy about the reasoner's response. The first is its recommendation to abandon the belief in conciliationism. And the second is its apparent incoherence: If one is to drop the belief that conciliationism is correct, why would one conciliate in response to the disagreement about the existence of free will? The main goal of the next chapter will be to put this feeling of uneasiness to rest. I think that the reasoner's recommendation is actually correct, at least for the context it's given. The problem is that the story recounted in Double Disagreement is underdescribed. In particular, I think that it is missing crucial information regarding the agent's relative degrees of confidence in the conclusions of its (first-order) reasoning about conciliationism, free will, and the existence of disagreements about these two matters. By embedding the scenario in $c_{12}$, we have implicitly filled in the missing

information in a particular way. Once we will have made this information explicit, the reasoner's recommendations will look much more plausible—or so, at least, I will argue.

# Chapter 6: Adding degrees of confidence

Let's start with two observations. First, a lot of the literature on peer disagreement formulates conciliatory views in terms of degrees of confidence, as opposed to categorical beliefs. What's more those pulled to conciliatory views typically put more trust in the more moderate versions of such views, according to which a disagreement with an epistemic peer should make one somewhat less confident that one's take on a complex issue is correct, as opposed to making one lower one's confidence dramatically. This is, of course, in stark contrast with the view implemented in our reasoner. Second, while we have been talking about Double Disagreement without mentioning the agent's (relative) degrees of confidence in her first-order reasoning about conciliationism, free will, and the existence of genuine disagreement about these two matters, it seems very intuitive that they should play some role in determining the correct doxastic response. For instance, if the agent's rational degree of confidence in the reasoning which has lead her to adopt conciliationism is lower than her rational degree of confidence in the reasoning which has lead her to conclude that the disagreement with Evelyn (regarding conciliatonism) is genuine, then, intuitively, she should give up her belief in conciliationism. If, on the other hand, the agent's rational degree of confidence in her reasoning about conciliationism is lower than her rational degree of confidence in her reasoning about the disagreement with Evelyn, then, intuitively, it's too much to require that she gives up her belief in conciliationism. And, of course, similar considerations apply to the agent's (relative)

degrees of confidence in her reasoning about free will and the disagreement about it.

What these observations suggest is that degrees of confidence may have an important role to play in our thinking about conciliatory views, as well as their behavior in Double Disagreement-like scenarios. If that's correct, then any model of conciliationism isn't fully adequate, as long as it doesn't take degrees of confidence into account. So we must add another twist to our model reasoner. As it turns out, enabling it to take into account (relative) degrees of confidence doesn't only shed light on the reasoner's surprising recommendation we discussed at the end of the last chapter, but also results in what I think is a fully adequate treatment of Double Disagreement-like scenarios.

To avoid a possible misunderstanding, let me emphasize that, throughout this chapter, when I talk about degrees of confidence what I have in mind are always *rational* degrees of confidence, or the degrees confidence that are justified in the agent's epistemic situation, or rational for the agent to have given her epistemic situation.[1] These shouldn't be confused with phenomenal feelings of confidence—although it's plausible to think that the two are related.

The remainder of this chapter is structured as follows. The more technical Sections 6.1–6.2 upgrade the model reasoner, drawing on the work of Pollock (1995, 2001, 2010). Section 6.3, then, returns to Double Disagreement, fills in the (missing) information about the agent's relative degrees of confidence in a couple different

[1]Cf. to Christensen's (2010b) "rational credences", Kelly's "reasonable credences", Lackey's (2010a, 2010b) "degrees of justified confidence", and Lasonen-Aarnio's (2013) "correct credences". Using some such notion as rational degree of confidence without providing an analysis of it is fairly standard in epistemology.

ways, shows how it can be captured formally, and explores the reasoner's responses to the scenario with this information. Finally, Section 6.4 discusses how the reasoner's recommendations correlate with the differences in the relative degrees of confidence, explains why this helps solve the problem of self-defeat, and contrasts the resulting solution with the other existing solutions from the literature.

## 6.1 Basic principles: Weakest Link and Winner Takes All

The input to our defeasible reasoner consists of statements of the form "My best first-order reasoning about whether $X$ suggests that $X$" from which it can defeasibly infer $X$. We have just noted that agents can be more or less confident in their first-order reasoning, and that this can affect how they should respond to the situation at hand. Consequently, we must assign different degrees of confidence, "strengths", or "weights" to the input states, depending on how confident the modeled agent is in her reasoning; and these strengths must be factored into computing the correct response to the context at hand.[2]

The reasoner we formulated in Chapter 5 determines the correct response to a context roughly as follows. It begins by constructing an argument framework, consisting of arguments and defeat relations among them, proceeds to select the winning arguments, and then outputs the conclusions supported by those arguments. But throughout this process, the reasoner takes all arguments to support their conclusions to the same degree. In order to factor degrees of confidence into the model, we will relativize the relation of support between arguments and conclusions to *degrees*

[2]Cf. (Pollock 1995, p. 101).

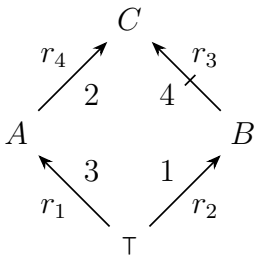



Figure 6.1: Weakest Link Principle

*of support*. Once we are done, it won't necessarily hold that any two arguments support their conclusions equally well, or to the same degree. Instead, what will typically happen is that one argument supports its conclusion to a greater degree, or more strongly, than the other argument supports its conclusion. The (relative) degrees of support will, then, have an effect on how conflicts between arguments are resolved and, thus, also on which arguments come out winning.

The idea that one (defeasible) argument can support its conclusion more strongly than another is both very natural and familiar.[3] My implementation of this idea in the model will draw on two basic principles that I take over from the work of Pollock.[4] I call these principles the *Weakest Link* and *Winner Takes All*.

The Weakest Link Principle, as its name suggest, says that an argument is only as good as its weakest element. A good way to get an immediate grasp on the principle is to see it at work. Consider the context $c_{15} = \langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W}$ empty and $\mathcal{R}$ comprised of the default rules $r_1 = \frac{\top}{A}$ , $r_2 = \frac{\top}{B}$ , $r_3 = \frac{B}{\neg C}$ , and $r_4 = \frac{A}{C}$ .[5] The corresponding graph is depicted in Figure 6.1. Notice the numbers next to the

[3]See e.g., (Dunne et al. 2011), (Grossi & Modgil 2015), (Modgil & Prakken 2013), (Pollock 2001, 2010), and (Prakken & Sartor 1997).

[4]See (Pollock 1995, 2001, 2010), especially, (Pollock 1995, Sec. 4.3)

[5]The questions about the reasonableness of the rules in $\mathcal{R}$ is orthogonal to the concerns of this section. So we can safely ignore the *Reasonable*-formulas and the reasonableness requirement.

arrows standing for rules. We use them to represent the relative strengths of rules. The fact that $r_2$ is associated with the number 1 doesn't mean anything by itself, but the fact that $r_2$ is associated with 1, while $r_3$ is associated with 4 means that $r_3$ is stronger, or that it has more weight, than $r_2$. So what numbers do for us, in effect, is linearly order the rules: $r_2$ (associated with 1) is the weakest default rule; it is followed by the slightly stronger $r_4$ (2); then comes the even stronger $r_1$ (3); and, finally, the rule $r_3$ (4) is the strongest.

Now let's compare the scenarios $\mathcal{S}_1 = \{r_1, r_2, r_4\}$ and $\mathcal{S}_2 = \{r_1, r_2, r_3\}$.[6] Notice that we have $\mathcal{W} \cup Conclusion[\mathcal{S}_1] \vdash C$ and $\mathcal{W} \cup Conclusion[\mathcal{S}_2] \vdash \neg C$. So the two scenarios are naturally thought of as arguments supporting opposing conclusions, and they both qualify as arguments in the technical sense of Section 5.1. Having attached numbers to rules, we can ask about the support that each of the two arguments confers on their respective conclusions. And this is where the Weakest Link Principle comes into play, saying that the support $\mathcal{S}_1$ confers on $C$ is 2, while the support $\mathcal{S}_2$ confers on $\neg C$ is 1. Why is that? Well, in the context of $\mathcal{S}_1$, we can reach $C$ by means of the chain $r_1$-$r_4$. This chain's weakest link is $r_4$ and its relative strength is 2. Similarly, in the context of $\mathcal{S}_2$, we can reach $\neg C$ by means of the chain $r_2$-$r_3$. This chain's weakest link is $r_2$ and its relative strength is 1. Thus, the case for $C$ is stronger than the case of $\neg C$, and in spite of the fact that the rule $r_3$ which directly supports $\neg C$ is stronger than the rule $r_4$ which directly supports $C$. It's worth emphasizing that the Weakest Link Principle is very general, letting one determine argument strengths in simple contexts like $c_{15}$, as well as much more

[6] If numbers are ignored, both $\mathcal{S}_1$ and $\mathcal{S}_2$ qualify as proper scenarios based on $c_{15}$.

complex ones.

But while the Weakest Link Principle lets us determine the relative strengths of arguments, it doesn't tell us how to handle conflicts between them. And we don't need to move past $c_{15}$ to see this: The argument $\mathcal{S}_1$ comes out stronger than the argument $\mathcal{S}_2$, and so, intuitively, $\mathcal{S}_1$ should win out. However, an important question remains: What is the overall or, what we might call, the *all-things-considered* degree of support of $C$, or its degree of support after all the relevant information? There are two candidate answers here, which seem to be equally plausible *prima facie*. First, we might say that the all-things-considered support of $C$ should equal the degree of support that the strongest argument for $C$ confers on it. This method for resolving conflicts is the principle I referred to as *Winner Takes All* above. And second, we might say that the degree of support that the strongest argument confers on $C$ should be taken as a starting point and that it should then be attenuated, in one way or another, by the argument for $\neg C$. Now, the Winner Takes All response is the one that Pollock gives, and it is the one that I will adopt here. This is not because I have a knockdown argument against the alternative, but because I only know how to capture the Winner Takes All formally.[7]

Note that we'll use the Winner Takes All method to resolve conflicting between

[7]There's formal work that could prove useful in capturing the alternative method for conflict resolution, namely, the recent and very interesting research on *numerical argumentation networks*—see e.g., (Barringer et al. 2012), (Gabbay 2012). Unfortunately, it is still in its infancy stage, and many complex issues need to be resolved before we could apply it in the present context. Chief among them is the questions of how to assign numerical weights when the underlying network—and you can just think of an inference graph here—contains cycles—see (Barringer et al. 2012, Sections 3–4). Pollock does offer an argument against the second response. His idea is that it commits one to the composition or accrual of reasons, or the view that two arguments for a conclusion can result in a higher degrees of support than either of the two arguments alone—we touched on this idea back in Section 2.1.3. Pollock thinks that we have independent reasons to reject this idea—see e.g., (Pollock 1995, pp. 101–4).

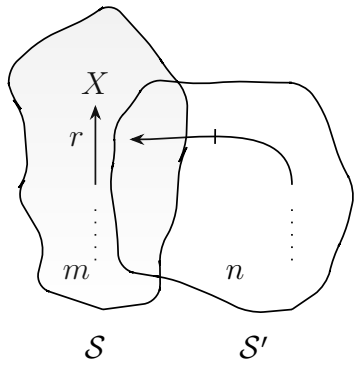



Figure 6.2: Winner Takes All for undermining defeat

arguments of the sort we just saw in $c_{15}$, as well when determining whether the support that some argument confers on a conclusion is excluded, or undermined, by another one. Let's say that we have two arguments $\mathcal{S}$ and $\mathcal{S}'$, that $\mathcal{S}$ supports some proposition $X$, and that $\mathcal{S}'$ contests the support that $\mathcal{S}$ confers on $X$ by supporting the proposition $Out(\mathfrak{r})$, where $\mathfrak{r}$ is the name of the rule $r$ that is one of the links in the chain of rules in $\mathcal{S}$ that leads to $X$. Let's also say that the relative strength with which $\mathcal{S}$ supports $X$ is $m$, while the strength with which $\mathcal{S}'$ supports $Out(\mathfrak{r})$ is $n$. (See Figure 6.2 for a schematic representation.) To apply the method here is to say that $\mathcal{S}'$ cancels all the support that $\mathcal{S}$ lends to $X$ if $n$ is greater or equal to $m$, and that $\mathcal{S}'$ has no effect on $\mathcal{S}$ otherwise.[8]

Now that we have a feeling for how the Weakest Link and Winner Takes All work, we can relativize support to degrees of support in the model.

[8]It's worth mentioning that there's an alternative to the Winner Takes All here: We could let the excluder do its work, no matter what its strength. Pollock considers the possibility and rejects it on the basis of its being "perverse"—see (Pollock 1995, pp. 103–4). For an argument to the opposite conclusion, see (Horty 2012, pp. 204–10).

## 6.2 Adding degrees to the model

To represent the information that a rule $r$ supports its conclusion to a greater degree, or more strongly, than another rule $r'$ we will use a device we have relied on in previous applications: a priority ordering on rules.[9] Let the statement $r \leq r'$ mean that the premise of the rule $r'$, $Premise[r']$, confers at least as much support on its conclusion, $Conclusion[r']$, as the premise of the rule $r$, $Premise[r]$, confers on its conclusion, $Conclusion[r]$. For the sake of simplicity, in what follows we will often drop the reference to rule premises and conclusions, reading $r \leq r'$ as saying that $r'$ is at least as strong or has at least as much weight as $r$. As in Parts I–II, we require that the relation $\leq$ satisfies the properties of reflexivity and transitivity.[10] But, in addition to this, here we will also require that $\leq$ satisfies the *connectivity* property,

$$r \leq r',\ r' \leq r, \text{ or both,}$$

according to which any two default rules are always comparable with respect to their strengths. Requiring that $\leq$ satisfies connectivity should make good sense, given what sort of information it's supposed to represent. For any two considerations conferring support to different conclusions, we'd seem to always be able to ask which of the

[9] When introducing the Weakest Link Principle above, I represented the information about the relevant strengths of rules using natural numbers. This was the easiest way to convey the basic intuition behind the principle. But one might wonder why I switch to a priority ordering here, instead of continuing to use natural numbers. The answer is that this is potentially misleading. Consider an abstract case that requires us to focus on three rules of increasing strength, $r$, $r'$, and $r''$, and that their relative strengths are represented using the numbers 1, 2, and 15. Quite naturally, one is led to think that $r''$ is much stronger than $r'$, while $r'$ is only a little bit stronger than $r$. So the use of numbers suggests that there's more structure beyond their ordinal ranking, or the fact $r''$ is stronger than $r'$ which, in turn, is stronger than $r$. However, in the formal model only ordinal ranking matters.

[10] Thus, for all rules $r$, $r'$, and $r''$, $r \leq r$, as well as if $r \leq r'$ and $r' \leq r''$, then $r \leq r''$.

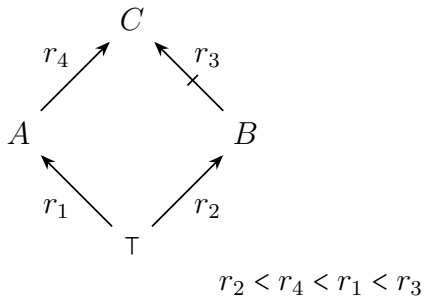



Figure 6.3: Weakest Link Principle, again

two confers more.[11] We will, again, make use of our shorthand: When we have $r \leq r'$ without $r' \leq r$, we will write $r < r'$. And when we have both $r \leq r'$ and $r' \leq r$ (for distinct rules), we will write $r \sim r'$.

In the remainder of this chapter, we'll be working with *weighted context*, or ordinary contexts extended with a priority relation on rules.

**Definition 6.1 (Weighted contexts)** *A* weighted context *$c$ is a structure of the form $\langle \mathcal{W}, \mathcal{R}, \leq \rangle$ where $\langle \mathcal{W}, \mathcal{R} \rangle$ is an ordinary context and $\leq$ is a reflexive, transitive, and connected relation (or a connected preorder) on $\mathcal{R}$.*

Recall the context $c_{15} = \langle \mathcal{W}, \mathcal{R} \rangle$. The information we previously captured using natural numbers can now be expressed using a preorder $\leq$ on $\mathcal{R}$. The result is the weighted context $c_{16} = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, where $\mathcal{W}$ and $\mathcal{R}$ are as before and $\leq$ is the ordering $r_2 < r_4 < r_1 < r_3$. The context is depicted in Figure 6.3.

With the information about the relative strengths of rules at our disposal, we can use it to compare scenarios, including those that qualify as arguments. Our next

[11] Cf. (Horty 2012, Section 1.1.2) who opts for a strict partial order, instead of a weak preorder, and explicitly rejects connectivity. But his context is different from ours, as Horty uses default logic to model reasons and reason interaction in different domains. Pollock (1994, 1995, 2001) sticks to the epistemic domain and requires connectivity, just as we do here.

definition will let us compare scenarios, according to the Weakest Link Principle:[12]

**Definition 6.2 (Weakest Link)** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a weighted context and $\mathcal{S}$ and $\mathcal{S}'$ two scenarios based on it. Then $\mathcal{S}'$ is at least as good as $\mathcal{S}$, written as $\mathcal{S} \preceq \mathcal{S}'$, if and only if there is a rule $r \in \mathcal{S}$ such that, for all $r' \in \mathcal{S}'$, $r \leq r'$.*

Let's apply the definition to $c_{16}$, starting with the arguments $\mathcal{S}_1 = \{r_1, r_4\}$ and $\mathcal{S}_2 = \{r_2, r_3\}$. It's not difficult to see that the latter argument $\mathcal{S}_2$ contains a rule, namely, $r_2$, that's weaker than both rules in $\mathcal{S}_1$—we have $r_2 < r_1$ and $r_2 < r_4$. And this means that $\mathcal{S}_1$ is at least as good as $\mathcal{S}_2$, or that $\mathcal{S}_2 \preceq \mathcal{S}_1$. Now, $\mathcal{S} \preceq \mathcal{S}'$, by itself, doesn't exclude the possibility that $\mathcal{S}' \preceq \mathcal{S}$ and that both arguments support their conclusions equally well. However, it's clear that in the particular case at hand $\mathcal{S}_1 \preceq \mathcal{S}_2$ doesn't hold. While one of the elements of $\mathcal{S}_2$, namely, the rule $r_3$, is stronger than both elements of $\mathcal{S}_1$ ($r_1 < r_3$ and $r_4 < r_3$), its other element, the rule $r_2$, is weaker than both of them ($r_2 < r_1$ and $r_2 < r_4$). Consequently, we have $\mathcal{S}_2 \prec \mathcal{S}_1$ and not $\mathcal{S}_1 \preceq \mathcal{S}_2$, or $\mathcal{S}_2 \prec \mathcal{S}_1$ in our shorthand.

Notice that Definition 6.2 can be used to compare any two scenarios. Our next example illustrates why this might seem problematic. When discussing *minimal* or *basic* arguments in Section 5.3, we considered the context $c_{14} = \langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W}$ consisting of formulas $A$ and $B$ and $\mathcal{R}$ consisting of the rules $r_4 = \frac{A}{C}$, $r_5 = \frac{B}{C}$, and $r_6 = \frac{C}{D}$.[13] We will now extend this context in three distinct ways. First, its

[12] The definition specifies how to lift a relation on rules to a relation on sets of rules. Not surprisingly, there are many alternative ways of lifting a relation on elements to sets of elements—see Barberà et al. (2004) for a thorough survey. The question of how our analysis might change if we opted for a different lifting procedure—perhaps, that of Brass (1991) or Horty (2012)—will have to be left for future work.

[13] We ignored the *Reasonable*-formulas then, and we will ignore them here as well. Nothing important hinges on this.

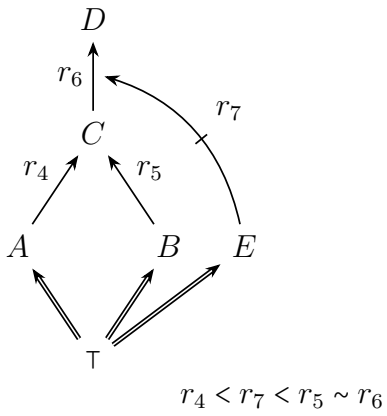



Figure 6.4: Multiple minimal arguments, extended

hard information $\mathcal{W}$ will now also include $E$. Second, its set of rules will now also include the rule $r_7 = \dfrac{E}{Out(\mathfrak{r}_6)}$. And third, there will now be a priority relation on the rules, namely, $r_4 < r_7 < r_5 \sim r_6$. The result is the weighted context $c_{17} = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, which is depicted in Figure 6.4. Now let's zoom in on the arguments $\mathcal{S}_3 = \{r_4, r_5, r_6\}$ and $\mathcal{S}_4 = \{r_7\}$. The argument $\mathcal{S}_3$ supports the conclusion $D$, while the argument $\mathcal{S}_4$ supports the conclusion $Out(\mathfrak{r}_6)$, challenging the support that $\mathcal{S}_3$ confers on $D$. If we compare $\mathcal{S}_3$ and $\mathcal{S}_4$ using Definition 6.2, we must conclude that $\mathcal{S}_4$ is a stronger argument than $\mathcal{S}_3$, or that $\mathcal{S}_3 \prec \mathcal{S}_4$. For $\mathcal{S}_3$ contains an element, the rule $r_4$, that's weaker than all the elements of $\mathcal{S}_4$. However, this result might seem counterintuitive: The argument $\mathcal{S}_3$ lets us reach $D$ without making use of $r_4$, that is, by means of the chain $r_5$-$r_6$. Thus, it might seem that our definition doesn't deliver the correct result when it ranks $\mathcal{S}_4$ higher than $\mathcal{S}_3$.

But the fault here does not actually lie with Definition 6.2, but, rather, with the fact that we are juxtaposing $\mathcal{S}_3$ and $\mathcal{S}_4$. The argument $\mathcal{S}_3$ is actually a combination of two basic arguments supporting $D$, namely, $\mathcal{S}_5 = \{r_4, r_6\}$ and $\mathcal{S}_6 = \{r_5, r_6\}$. When

we juxtapose $\mathcal{S}_5$ and $\mathcal{S}_6$ with $\mathcal{S}_4$, we get the intuitive result, or the result that one would expect the Weakest Link Principle to deliver, $\mathcal{S}_5 \prec \mathcal{S}_4$ and $\mathcal{S}_4 \prec \mathcal{S}_6$. The upshot is simple: It's important that the definition capturing the principle gets applied to minimal arguments. This is why it's more natural to use the notion of *basic defeat* as our springboard for specifying the notion of defeat that takes into account varying degrees of support. Just like we did in Section 5.3, we will first specify the conditions under which one minimal argument defeats another and then extrapolate the relation of defeat to arguments of arbitrary complexity:

**Definition 6.3 (Basic defeat, with the weakest link)** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a weighted context and $\mathcal{S}$ and $\mathcal{S}'$ two arguments from the argument framework $\mathcal{F}(c)$ based on it. Then $\mathcal{S}$* basic-defeats *$\mathcal{S}'$, written as $\mathcal{S} \rightsquigarrow_b^w \mathcal{S}'$, if and only if there is some rule $r$ in $\mathcal{S}'$ such that*

*(i) $\mathcal{S}'$ is in $Minimal_{\mathcal{F}(c)}(r)$ and*

*(ii) either*

*(1) $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$ and $\mathcal{S}$ is in $Minimal_{\mathcal{F}(c)}(\neg Conclusion[r])$,*

*or*

*(2) $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(\mathfrak{r})$ and $S$ is in $Minimal_{\mathcal{F}(c)}(Out(\mathfrak{r}))$,*

*and $\mathcal{S}' \preceq \mathcal{S}$.*

The only difference between this definition and Definition 5.11 (Basic defeat) from Section 5.3 is the additional requirement that the defeating argument is at least as

strong as the defeated one. An easy check suffices to see that we get $\mathcal{S}_4 \rightsquigarrow_b^w \mathcal{S}_5$ and $\mathcal{S}_6 \rightsquigarrow_b^w \mathcal{S}_4$, while we do not get $\mathcal{S}_4 \rightsquigarrow_b^w \mathcal{S}_3$. With this, we are done. Having incorporated the Weakest Link Principle into the notion of basic defeat, we can simply reuse the definition of (nonbasic) defeat from Section 5.3:

**Definition 6.4 (Defeat, with the weakest link)** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ and $\mathcal{S}'$ two arguments based on it. Then $\mathcal{S}$ weakest-link-defeats $\mathcal{S}'$, $\mathcal{S} \rightsquigarrow^w \mathcal{S}'$, if and only if there is an $\mathcal{S}'' \subseteq \mathcal{S}$ and an $\mathcal{S}''' \subseteq \mathcal{S}'$ such that $\mathcal{S}'' \rightsquigarrow_b^w \mathcal{S}'''$.*

Our next definition specifies how to construct argument frameworks from weighted contexts. Not surprisingly, the only difference is that now we rely on the weakest-link-defeat.

**Definition 6.5 (Argument frameworks for weighted context)** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a weighted context. Then the* argument framework $\mathcal{F}(c)$ *based on $c$ is the pair $\langle \mathcal{A}, \rightsquigarrow^w \rangle$ where $\mathcal{A}$ is the set $Arguments(c)$ and $\rightsquigarrow^w$ is the set $\{(\mathcal{S}, \mathcal{S}') \in \mathcal{A} \times \mathcal{A} : \mathcal{S}$ weakest-link-defeats $\mathcal{S}'\}$.*

We can apply stability and preference semantics to frameworks built from weighted contexts just as we applied them to frameworks built from ordinary ones. So nothing needs to be changed here. The observation with the statement of which we close this section makes it plain that the addition of weights to contexts is a conservative extension of our original framework—the proof is in the Appendix:

**Observation 10** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be an ordinary context and $c' = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be the same context with a connected preorder $\leq$ assigning all the rules $r$ in $\mathcal{R}$ the same*

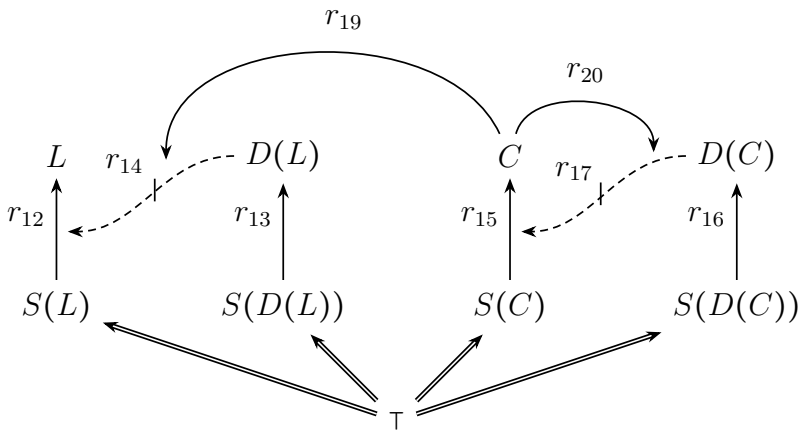



Figure 6.5: Double Disagreement, once more

*weight—so, for all* $r$, $r' \in \mathcal{R}$, $r \sim r'$. *Then* $\mathcal{F}(c) = \mathcal{F}(c')$.

## 6.3 Double Disagreement with degrees

Now let's see how the tools developed in Sections 6.1–6.2 can help us with Double Disagreement, which we captured by the context $c_{12} = \langle \mathcal{W}, \mathcal{R} \rangle$, depicted graphically here one final time—see Figure 6.5. At the end of Chapter 5, we saw that our (sophisticated) defeasible reasoner doesn't conclude either $C$ or $L$ from $c_{12}$, suggesting that the correct response to Double Disagreement is to abandon one's belief in conciliationism, as well as one's belief in the existence of free will. We also noted that there appears to be something incoherent about this response: Why would the agent conciliate in response to the disagreement about libertarian free will if it is to abandon the belief in concliationism? I suggested that we find this response counterintuitive because the scenario is underdescribed. I also promised that the reasoner's recommendations will look more compelling, once we supply the original description of Double Disagreement with the information missing from it

and enable the reasoner to take it into account.

The missing information, of course, concerns the reasoning agent's relative degrees of confidence in her first-order reasoning about conciliatonism, free will, and the disagreements about these two matters, expressed by the propositions $Seems(C)$, $Seems(L)$, $Seems(Disagree(C))$, and $Seems(Disagree(L))$. It seems intuitive to expect that these degrees of confidence, and they alone, would determine the relative degrees of support of the arguments that can be constructed from $c_{12}$. But how can we make this work? How are we to translate degrees of confidence into degrees of support? As a first step toward answering these questions, note that $c_{12}$ contains three types of rules. To be more precise, there are rules of the form $r(X) = \dfrac{Seems(X)}{X}$, rules of the form $r'(X) = \dfrac{Disagree(X)}{Out(\mathfrak{r}(X))}$, as well as those of the form $\dfrac{C}{Reasonable(\mathfrak{r}'(X))}$. Our task is to express the degrees of confidence as an ordering on these different types of rules. One thing appears to be clear: The relative degrees of confidence must correspond directly to the relative weights of the first type of rules. Thus, if we want to capture the version of Double Disagreement where the agent is more confident of its reasoning about conciliationism than its reasoning about free will, we would assign $r_{15}$ more weight than $r_{12}$, or set $r_{12} < r_{15}$.

But what about the remaining two types of rules? Notice that due to the structure of the context $c_{12}$, all arguments in $\mathcal{F}(c_{12})$ that contain a rule of the form $\dfrac{C}{Reasonable(\mathfrak{r}'(X))}$ must also contain the rule $r_{15} = \dfrac{Seems(C)}{C}$ and all arguments that contain a rule of the form $\dfrac{Disagree(X)}{Out(\mathfrak{r}(X))}$ must also contain a rule of the form $\dfrac{Seems(Disagree(X))}{Disagree(X)}$, as well as the rule $r_{15}$. Why is this important? Well, since our reasoner relies on the Weakest Link Principle to determine argument

strength, the arguments of $\mathcal{F}(c_{12})$ can only ever be as strong as the rules of the form $\dfrac{Seems(X)}{X}$ that they contain. So there's a clear sense in which what really matters for the relative degrees of support of the arguments and, thus, also for the overall conclusions the reasoner draws are the relative weights of the *four* rules that have this form, namely, $r_{12}$, $r_{13}$, $r_{15}$, and $r_{16}$.

This is almost all that matters, but not quite. The calculations of support that the arguments from $\mathcal{F}(c_{12})$ confer on their conclusions will depend on one further assumption that I want to be explicit about. We can call it *No Support Lost*. For starters, consider the argument $\mathcal{S}_7$ comprised of the rules $r_{15} = \dfrac{Seems(C)}{C}$ and $r_{19} = \dfrac{C}{Reasonable(\mathfrak{r}_{14})}$. Due to the Weakest Link Principle, $\mathcal{S}_7$ can support the formula $Reasonable(\mathfrak{r}_{14})$ only as strongly as $r_{15}$ supports the conclusion $C$. However, it's in principle possible that the support conferred on $Reasonable(\mathfrak{r}_{14})$ is much weaker than the support conferred on $C$. This would happen in all the cases where the weight of $r_{19}$ is much lower than that of $r_{15}$. These are exactly the sorts of cases that our assumption will rule out.

**No Support Lost**:

The weights of the rules of the form $\dfrac{Disagree(X)}{Out(\mathfrak{r}(X))}$ and $\dfrac{C}{Reasonable(\mathfrak{r}'(X))}$ must be at least as high as the weights of the rules of the form $\dfrac{Seems(X)}{X}$ that they depend on.

The intuitive notion of dependency between rules that this assumption appeals to can be made precise in terms of the relations between the arguments in $\mathcal{F}(c_{12})$: If there's no argument $\mathcal{S}$ in $\mathcal{F}(c_{12})$ that contains $r$ but not $r'$, then $r$ depends on $r'$. The

No Support Lost assumptions strikes me as very intuitive. Whats' more, nothing of importance appears to hinge on us making it.[14] Having stated the assumption explicitly, we can turn to the question of how the reasoner's recommendations depend on the relative weights of the four crucial rules. I will provide a general answer to this question at the beginning of the next section. The remainder of this section will explore a few ways of extending $c_{12}$ with specific orderings on rules, expressing various versions of Double Disagreement. It can be skipped without loss of continuity.

We begin with a couple of versions of Double Disagremeent where the agent's degree of confidence in her reasoning about conciliationism is higher than her degree of confidence in there being a genuine disagreement about it. The first ordering we'll be looking at is $\leq_1$.

$\leq_1$: $r_{12} < r_{14} \sim r_{13} < r_{16} \sim r_{17} < r_{15} \sim r_{19} \sim r_{20}$

This particular ordering represents a version of Double Disagreement where the reasoning agent is most confident in $Seems(C)$, followed by $Seems(Disagree(C))$, then by $Seems(L)$, and least confident in $Seems(Disagree(L))$. Notice that this ordering assigns the rules of the form $\dfrac{Disagree(X)}{Out(\mathfrak{r}(X))}$ and $\dfrac{C}{Reasonable(\mathfrak{r}'(X))}$ the same relative weight as the rules they depend on—or, more precisely, the weakest of rules of the form $\dfrac{Seems(X)}{X}$ they depend on. This is only for the sake of simplicity. Nothing would change if these rules were assigned different relative weights, at least as long as No Support Lost wasn't violated.

Extending $c_{12} = \langle \mathcal{W}, \mathcal{R} \rangle$ with $\leq_1$ results in the weighted context $c_{18} = \langle \mathcal{W}, \mathcal{R}, \leq_1 \rangle$

[14] Without the assumption, things are messier. The question of how dropping the assumption might affect the analysis is left for future work.

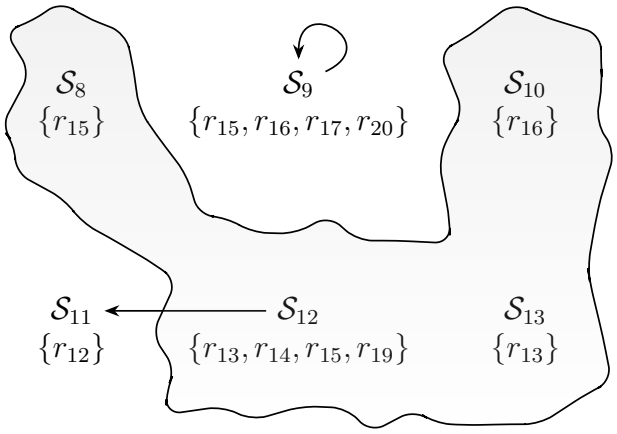



Figure 6.6: Core arguments from $\mathcal{F}(c_{18})$, given $\leq_1$

and the corresponding argument framework $\mathcal{F}(c_{18})$. We'll be looking at a fragment of this framework that contains its most informative minimal arguments, namely:

the $C$-minimal argument $\mathcal{S}_8 = \{r_{15}\}$;

the $Out(\mathfrak{r}_{15})$-minimal $\mathcal{S}_9 = \{r_{15}, r_{16}, r_{17}, r_{20}\}$;

the $Disagree(C)$-minimal $\mathcal{S}_{10} = \{r_{16}\}$;

the $L$-minimal $\mathcal{S}_{11} = \{r_{12}\}$;

the $Out_{12}$-minimal $\mathcal{S}_{12} = \{r_{13}, r_{14}, r_{15}, r_{19}\}$; and

the $Disagree(L)$-minimal argument $\mathcal{S}_{13} = \{r_{13}\}$.

The fragment is depicted graphically in Figure 6.6. Notice that there are fewer defeat relations among argument, when compared to the corresponding fragment of the framework $\mathcal{F}(c_{12})$ constructed from $c_{12}$. Let's start by zooming in on the $C$-minimal argument $\mathcal{S}_8$ and the $Out(\mathfrak{r}_{15})$-minimal argument $\mathcal{S}_9$. In $\mathcal{F}(c_{12})$, $\mathcal{S}_9$ defeated $\mathcal{S}_8$, but

it doesn't do so here. The reason is that the crucial condition of the Definition 6.4 (Section 6.2) is not satisfied: It's simply not the case that $\mathcal{S}_8 \preceq \mathcal{S}_9$. For $\mathcal{S}_8 \preceq \mathcal{S}_9$ to obtain, there would have to be a rule in $\mathcal{S}_8$ that's weaker than every rule in $\mathcal{S}_9$. But $\mathcal{S}_8$ contains only one rule, $r_{15}$, and $\mathcal{S}_9$ happens to include $r_{16}$ which is weaker than $r_{15}$. As a result, $\mathcal{S}_9$ doesn't defeat $\mathcal{S}_8$. Similarly, $\mathcal{S}_9$ no longer defeats $\mathcal{S}_{12}$. The other defeat relations remain. First, given that $\mathcal{S}_9 \preceq \mathcal{S}_9$, the argument $\mathcal{S}_9$ still self-defeats. And second, the $Out(\mathfrak{r}_{12})$-minimal argument $\mathcal{S}_{12}$ still defeats the $L$-minimal $\mathcal{S}_{11}$. It's easy to see that the rule $r_{12}$ from $\mathcal{S}_{11}$ is weaker than all of the rules from $\mathcal{S}_{12}$, or that $\mathcal{S}_{11} \preceq \mathcal{S}_{12}$.

Since both $\mathcal{S}_8$ and $\mathcal{S}_{12}$ are no longer defeated by $\mathcal{S}_9$, they get included in the preferred extension of $\mathcal{F}(c_{18})$, as the picture illustrates. Further, given that $\mathcal{S}_8$ is in the preferred extension, while $\mathcal{S}_{11}$ is not, the formula $C$ follows from $c_{18}$, while $L$ does not. Advocates of conciliatory views should be pleased with this result: In spite of the disagreement about conciliationism, our reasoner suggests that the correct doxastic response is to stick to one's belief in conciliatonism and to abandon one's belief in libertarian free will.

One might expect that any ordering assigning the rule $r_{15}$ more weight than the rule $r_{16}$ will lead to this result. But the next ordering $\leq_2$ demonstrates that this doesn't hold true.

$\leq_2$: $r_{16} \sim r_{17} < r_{13} \sim r_{14} < r_{12} < r_{15} \sim r_{19} \sim r_{20}$

The ordering $\leq_2$ encodes a version of Double Disagreement where the agent is most confident in $Seems(C)$, followed by $Seems(L)$, then by $Seems(Disagree(L))$, and

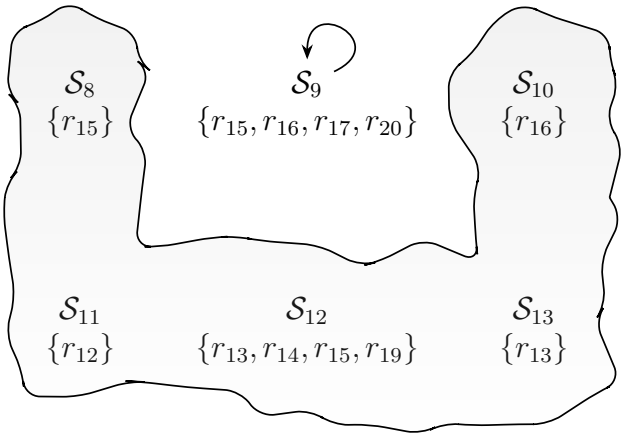



Figure 6.7: Core arguments from $\mathcal{F}(c_{19})$, given $\leq_2$

least confident in $Seems(Disagree(C))$.

Let $c_{19} = \langle \mathcal{W}, \mathcal{R}, \leq_2 \rangle$ be the result of extending $c_{12}$ with this ordering. The relevant fragment of the argument framework $\mathcal{F}(c_{19})$ is depicted in Figure 6.7. Similarly to what we saw in the case of $\mathcal{F}(c_{18})$, the argument $\mathcal{S}_9$ self-defeats without defeating either $\mathcal{S}_8$ or $\mathcal{S}_{12}$. What's different here is the relation between the $L$-minimal $\mathcal{S}_{11}$ and the $Out(\mathfrak{r}_{12})$-minimal $\mathcal{S}_{12}$. Since $\mathcal{S}_{11}$ consists of only one element $r_{12}$ that's stronger than some of the element of $\mathcal{S}_{12}$, we don't get $\mathcal{S}_{11} \preceq \mathcal{S}_{12}$, meaning that $\mathcal{S}_{12}$ doesn't defeat $\mathcal{S}_{11}$. Since both $\mathcal{S}_8$ and $\mathcal{S}_{11}$ are in the preferred extension of $\mathcal{F}(c_{19})$, both $C$ and $L$ follow from the context $c_{19}$. Thus, our reasoner suggests that the correct response to the case at hand is to stick to one's belief in conciliationism, as well as to one's belief in free will. At first blush, this recommendation might look incoherent: Why would an agent who believes that conciliationism is correct respond to the disagreement about $L$ like a steadfaster? But we shouldn't forget that $c_{19}$ captures a version of Double Disagreement where the agent is more confident of its reasoning about

free will than its reasoning about the existence of a genuine disagreement about it. Given this, the reasoner's recommendation actually seems perfectly sensible. In the end, we can easily imagine the reasoning agent going through the following line of thought: I do believe that conciliationism is correct. However, that doesn't mean that I should abandon my beliefs in response to every disagreement. I would abandon my belief in free will in response to my disagreement with Milo if I was at least as confident in it being a genuine disagreement as I am of my reasoning about free will. But I'm not.

An even more surprising result is revealed by the third ordering we'll be looking at:

$$\leq_3\colon\ r_{16} \sim r_{17} < r_{14} \sim r_{15} \sim r_{19} \sim r_{20} < r_{12} < r_{13}$$

Extending $c_{12}$ with $\leq_3$ gives us the weighted context $c_{20} = \langle \mathcal{W}, \mathcal{R}, \leq_3 \rangle$. The argument framework $\mathcal{F}(c_{20})$ constructed from it happens to coincide with the framework $\mathcal{F}(c_{19})$, the fragment of which we have just looked at. The preferred extensions of $\mathcal{F}(c_{20})$ and $\mathcal{F}(c_{19})$ are also the same. So our reasoner's recommendation for $c_{20}$ is, again, to hold onto the belief in conciliationism, as well as to the belief in the existence of free will. But, contrary to $\leq_2$, the ordering $\leq_3$ assigns more weight to $r_{12}$ than it does to $r_{13}$, meaning that we're dealing with a version of the scenario where the agent is less confident in its reasoning about libertarian fee will than there being a genuine disagreement about it. So one may, again, feel that the recommendation is incoherent. However, it too turns out to be perfectly sensible: Since $\leq_3$ assigns more weight to $r_{12}$ than it does to $r_{15}$, the scenario is one where the agent is more

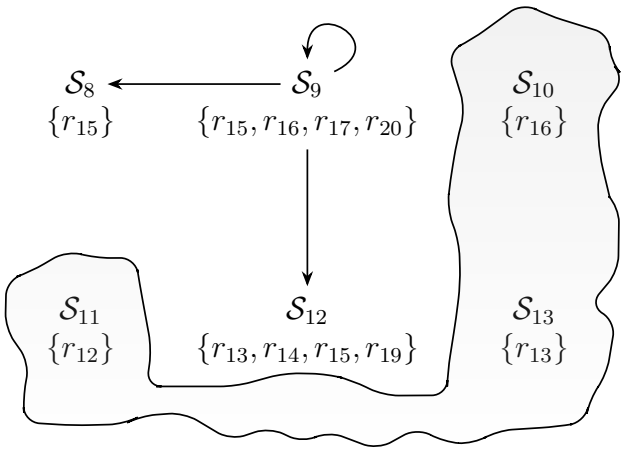



Figure 6.8: Core arguments from $\mathcal{F}(c_{21})$, given $\leq_4$

confident in its reasoning about free will than its reasoning about conciliationism. And we can easily imagine her thinking as follows: After thinking about the epistemic significance of disagreement, conciliationism seems to me to be correct. If it is, I should abandon my belief in the existence of free will in response to my disagreement with Milo. However, I am more confident of my reasoning about free will than my reasoning about conciliationism, and it would be foolish to distrust the conclusion of my reasoning about free will on the basis of a view I am less confident in than this reasoning.

So much for the versions of Double Disagreement where the agent is more confident in conciliationism than there being a genuine disagreement about it. The final two orderings we're going to explore model cases where the agent has more confidence in the disagreement about conciliationism being genuine than conciliationism itself. The first of them is $\leq_4$.

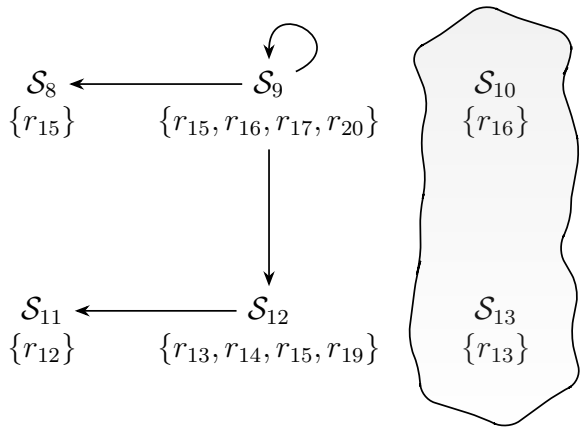
$\leq_4$: $r_{14} \sim r_{15} \sim r_{17} \sim r_{19} \sim r_{20} < r_{12} < r_{13} < r_{16}$

Once we add $\leq_4$ to $c_{12}$, we acquire the weighted context $c_{21} = \langle \mathcal{W}, \mathcal{R}, \leq_4 \rangle$, with the corresponding argument framework $\mathcal{F}(c_{21})$. Figure 6.8 depicts the relevant fragment of this framework. It's not difficult to see that here we have $\mathcal{S}_8 \preceq \mathcal{S}_9$. So the defeat relation between $\mathcal{S}_9$ and $\mathcal{S}_8$ stays in place, as do the relations between $\mathcal{S}_9$ and itself and $\mathcal{S}_9$ and $\mathcal{S}_{12}$. Now consider the $L$-minimal $\mathcal{S}_{11}$ and its potential defeater $\mathcal{S}_{12}$. Since $\mathcal{S}_{12}$ contains elements that are weaker than $r_{12}$, $\mathcal{S}_{12}$ does not defeat $r_{11}$. The preferred extension of $\mathcal{F}(c_{21})$ includes $\mathcal{S}_{11}$, but does not include $\mathcal{S}_8$. Therefore, $L$ does, but $C$ does not follow from $c_{21}$.

The reasoner's recommendation to the particular version of Double Disagreement modeled is both intuitive and well in line with the ideas about the behavior of conciliatory views in the literature—more on this in the next section. As it turns out, however, not every ordering that assigns more weight to $r_{16}$ than to $r_{15}$ lead to it. This is witnessed by $\leq_5$.

$\leq_5$: $r_{12} < r_{13} \sim r_{14} < r_{15} \sim r_{17} \sim r_{19} \sim r_{20} < r_{16}$

Extending $c_{12}$ with $\leq_5$ results in the weighted context $c_{22} = \langle \mathcal{W}, \mathcal{R}, \leq_5 \rangle$. The relevant fragment of $\mathcal{F}(c_{22})$ is depicted in Figure 6.9. Notice that this is the same graph that we saw in Section 5.4 before we turned to degrees of confidence. The addition of $\leq_5$ to $c_{12}$ has, thus, left the defeat relations between arguments intact. Since neither $\mathcal{S}_8$, nor $\mathcal{S}_{11}$ are in the preferred extension of $\mathcal{F}(c_{22})$, the reasoner's recommended response is to abandon both one's belief in conciliationism and one's belief in free will. Are we, then, back to where we started? Clearly not. For, having taken the degrees of confidence into account, we can see that the reasoner doesn't issue this seemingly




$$r_{12} < r_{13} \sim r_{14} < r_{15} \sim r_{17} \sim r_{19} \sim r_{20} < r_{16}$$

Figure 6.9: Core arguments from $\mathcal{F}(c_{22})$, given $\leq_5$

incoherent recommendation in all versions of Double Disagreement. What's more, now we also have a better grasp on the conditions under which it does. The ordering $\leq_5$ expresses a situation where the agent's degree of confidence in her reasoning about the existence of free will is lower than *both* her degree of confidence in her first-order reasoning about conciliationism and her degree of confidence in there being a genuine disagreement about free will. Is it really incoherent for the agent to abandon both of these belief if those are her relative degrees of confidence? I don't think it is, but I will defer the explanation to the next section.

## 6.4 Discussion

Now that we have seen how the agent's relative degrees of confidence get captured by an ordering on the rules of $c_{12}$ and that we have looked at a few examples, we can give a fully general answer to the following question: How do the reasoner's recommendations depend on these degrees of confidence? Or, to put the

same question in different terms, what are the reasoner's recommendations for all the different versions of Double Disagreement? We'll start with the recommendations regarding the belief in conciliationism, and then proceed to the recommendations regarding free will—which is, of course, a proxy for recommendations regarding any other theoretical question. It'll be useful to introduce some simple formal notation here: Recall that $Seems(X)$ expresses the proposition that the agent has reasoned about whether $X$ to the best of her ability, arriving at the conclusion that $X$. Now, $Seems(X) \leq Seems(Y)$ will express the idea that the agent is at least as confident in the conclusion of her reasoning about whether $X$ as she is about the conclusion of her reasoning about whether $Y$; and $Seems(X) < Seems(Y)$ will be a shorthand for $Seems(X) \leq Seems(Y)$ and not $Seems(Y) \leq Seems(X)$.

The recommendation regarding the belief in conciliationism depend only on the relative degrees of confidence in $Seems(C)$ and $Seems(Disagree(C))$. The reasoner suggests that one is to abandon one's belief in conciliationism, if one's degree of confidence in one's (first-order) reasoning about conciliationism is only as high as one's degree of confidence in there being a genuine disagreement about it—or, formally, if $Seems(C) \leq Seems(Disagree(C))$—and that one is to retain this belief otherwise—that is, if $Seems(C) > Seems(Disagree(C))$. These recommendation appear to be perfectly intuitive. Turning to the recommendations regarding free will, what matters here are the relative degrees of confidence in $Seems(C)$, $Seems(L)$, and $Seems(Disagree(L))$. The reasoner suggests that one is to abandon one's belief in the existence of free will, in case one's degree of confidence in one's (first-order) reasoning about it is only as high as one's degree of confi-

dence in there being a genuine disagreement about it *and* one's degree of confidence in one's (first-order) reasoning about conciliationism; and that one is to retain this belief otherwise. Or, more formally, one is to abandon the belief in $L$ if $Seems(L) \leq Seems(Disagree(L))$ and $Seems(L) \leq Seems(C)$; and one is to retain it if either $Seems(L) > Seems(Disagree(L))$ or $Seems(L) > Seems(C)$.

The reasoner's recommendations have two surprising features. The first is that the question of whether one is to retain the belief in free will turns out to depend on the relative degrees of confidence in one's first-order reasoning about free will and conciliationism. And the second surprising feature is that this question does *not* depend on whether or not conciliationism turns on itself, or whether or not one is to abandon one's belief in conciliationism. Why are these two features surprising? Well, because they are in stark contrast with the majority view on the behavior of conciliatory views in the literature.

Let's start with the first feature—which is actually independent of the worries about the behavior of conciliatonism in Double Disagreement-like scenarios. The proponents of conciliatory views readily take on board the idea that the correct doxastic response to a typical disagreement situation—a conciliatory agent finds herself disagreeing about free will with an epistemic peer—will depend on the agent's (initial) degrees of confidence in free will: The higher the agent's pre-disagreement degree of confidence in free will is, the higher her rational post-disagreement degree of confidence will be. Some proponents of conciliatory views might also accept the idea that the correct doxastic response to a typical disagreement scenario will depend on the agent's pre-disagreement degree of belief in conciliationism: The more

confident the agent is in the correctness of conciliationism, the lower her rational post-disagreement degree of confidence in the existence of free will will be. But even though the proponents of conciliatory views think that the degrees of confidence play a role in determining the correct doxastic responses, they also proceed under the assumption that the roles of these degrees are independent of each other: In particular, they largely take it for granted that the agent is to lower her confidence in the existence of free will, no matter how it compares to her confidence in conciliationism.[15] We have seen already that it's not difficult to make intuitive sense of an agent who retains her belief in free will, in spite of finding herself in a disagreement about free will and believing that conciliationism is correct: She's just more confident of her (first-order) reasoning about free will than her (first-order) considerations about conciliationism. This prompts the question of why does the peer disagreement literature assume the contrary. I suspect that this due to the fact that it tends to operate with informal models of conciliatory views that do not distinguish between the agent's doxastic attitudes and the domain-specific reasoning that leads to them.

And now for the second surprising feature and the literature focusing specifically on the problem of self-defeat. The majority view among those thinking that the problem has a solution can be stated in one sentence: If the correct doxastic

[15] Although this claim is correct, we might need to add some further qualifications to it to make it fit all the various views and ideas from the literature. For instance, at least some conciliationists, including Christensen (2009) and Matheson (2015a,c), would acknowledge the existence of correlations between the degrees of confidence in conciliatonism and free will. Matheson, in particular, would say that the agent should retain full confidence in free will in any case where she has overwhelming (misleading) evidence that conciliationism is false. Still, this doesn't change the fact that he proceeds under the assumption that the relative degrees of confidence in free will and conciliationism play no role in the typical scenarios.

response to at least some Double Disagreement-like scenarios involves abandoning one's belief in concliationism, then it also involves retaining one's belief in the existence of libertarian free will. The proposed solutions to the problem, then, fall into two camps: those that deny the antecedent of this sentence, and those that concede that it is true. The first camp comprises the solutions of Tomas Bogardus (2009), Christensen (2013), Elga (2010), and John Pittard (2015). Bogardus argues that we have a special rational insight into the truth of conciliationism which makes it immune to any disagreement about it.[16] Christensen admits that conciliationism can turn on itself, and that it does turn on itself in Double Disagreement-like scenarios. However, he also argues that this doesn't mean that one is to abandon one's belief in conciliationism. What Disagreement-like scenarios show instead is that there are inherently unfortunate or tragic epistemic situations—of which they are but one type of example—such that even the optimal response to them involves a violation of some *rationality ideal*.[17] But, still, the optimal response involves retaining the belief in conciliationism and lowering the degree of confidence in free will. Elga, in turn, takes the self-defeat challenge to be fatal for standard conciliatory views, but he also goes on to defend self-exempting *partially conciliatory views* that demand conciliation in response to any disagreement, except for disagreements about such views themselves.[18] Finally, Pittard argues that all types of conciliatory views self-

[16]See (Bogardus 2009, especially, pp. 332–3).

[17]An agent who retains her belief in conciliationism in a Double Disagreement scenario, in particular, is bound to violate either the ideal of *Respecting evidence of error*—by not forming the belief that her level of confidence in conciliationism is too high—or the ideal of *Level-connection*—by forming this belief, but retaining full confidence in conciliationism. See (Christensen 2013, especially, pp. 90–6).

[18]See (Elga 2010, especially, Sections 7–8).

exempt—and not only partially conciliatory ones, as suggested by Elga—because they do not issues a clear recommendation for how to respond to a disagreement about conciliationism, in spite of all the seemings to the contrary.[19]

The second camp comprises the solution of Jonathan Matheson (2015a,b,c), and, thus, may appear to be less well-represented. However, I suspect that most advocates of conciliatory views assume that some solution that goes along the lines of Matheson's has to work. Now, Matheson suggests that Double Disagreement-like scenarios come in two kinds: those where conciliatory views do not turn on themselves and those where they do. In the former type of scenarios, the agent is to retain the belief in conciliationism and to lower her confidence in the existence of libertarian free will. In the latter type of scenarios, the agent is to abandon the belief in conciliationism and to retain the belief in free will. While Matheson does admit that, in the second type of cases, conciliationism actually recommends abandoning both itself and the belief in the libertarian free will, he also think that the agent is to follow only the first recommendation. Why? Well, Matheson suggests that the first recommendation is *higher-order*—or a recommendation for which recommendations to follow—and that it, therefore, has the ability to undercut the second one. What's more, he also thinks that the agent is to follow her evidence in a

[19]Here's Pittard's (2015) idea in a little more detail: He starts off suggesting that what's basic to all sorts of conciliatory views is a commitment to showing *epistemic deference*. So, when a conciliationist disagrees with someone, they are to exhibit deference to this person. In a standard case of disagreement over some neutral proposition $Q$ the deferential response is clear: A conciliationist is to reduce her confidence in $Q$. A similar response may seem appropriate in Double Disagreement-like scenarios too, but here Pittard bring in the distinction between two dimensions or "levels" at which one's response to a disagreement can be deferential—the *belief level* and the *reasoning level*—and argues that no response to a disagreement about the correct way to disagree can show deference at both levels. If a conciliationist tries to be deferential at the level of belief, she is being nondeferential at the level of reasoning, and vice versa.

"downstream direction" beginning by following the higher-order recommendations and then proceeding to the lower-order ones.[20] So an agent who follows her evidence in an unfortunate scenario will follow only the higher-order recommendation of conciliationism and never encounter the lower-order one, ending up abandoning the belief in conciliationism and sticking to the belief in the existence of free will.[21]

So the majority view in the literature is that the seemingly incoherent doxastic response—believing neither conciliationism, nor free will—is never correct. The one exception to it is Clayton Littlejohn (2019) who, at least, entertains the thought that there may be nothing wrong with such responses. This is illustrated by the following passage:

> [..A] conciliatory thinker can continue to suspend when peers disagree without having any attitudes at all towards [conciliationism]. The two things recommended (i.e., being conciliatory on some contested propositions, suspending on [conciliationism]) are perfectly possible to do together. Thus, they aren't incompatible. The [self-defeat] objections simply misses its intended target (Littlejohn 2019, p. 4).[22]

Since Littlejohn's defense of these responses is confined to drawing an analogy with

[20]This is supposed to follow from evidentialism, or, roughly, the view centered around the evidentialist requirement we discussed in Chapter 3.

[21]See (Matheson 2015a, especially, Section 4) and (Matheson 2015c, pp. 153–7).

[22]The text actually says, "The simple [self-defeat] objection simply misses its intended target." Littlejohn goes on to discuss the "subtle" objection, as well as to provide a response to it. His discussion makes it clear that the subtle objection depends on two principles, and that any plausible view on rational belief and its connection to first- and higher-order evidence will accept at most one of them. In light of the way we have been thinking about conciliationism, the principle Littlejohn dubs *rational conversion* would seem to be particularly suspect. But either way, the subtle objection is not something that we need to worry about here.

the practical domain, I think he is best read as appealing to a burden of proof here.[23] As long as the opponents of conciliatory views haven't explained what's wrong with the apparently incoherent responses, the advocates of the views don't even have anything to worry about.

But let's leave considerations about burdens of proof aside and take a closer look at those versions of Double Disagreement for which the reasoner recommends the seemingly incoherent response. In a typical scenario where this happens, the agent's degree of confidence in her (first-order) reasoning about free will will be lower than her degree of confidence in conciliationism, as well as her degree of confidence in her disagreement with Milo—that is, the disagreement regarding free will—being genuine.[24] What's more, her degree of confidence in her reasoning about conciliationism can only be as high as her degree of confidence in there being a genuine disagreement about it. Now let's try putting ourselves into the agent's shoes. It seems easy enough to imagine her entertaining the following train of thought: I've thought about the epistemic significance of disagreement to the best of my ability, and, as far as I can tell, conciliationism is correct. If it is indeed correct, I shouldn't trust my reasoning about free will—as my disagreement with Milo provides a good reason to think that my reasoning about it might rest on a mistake. I'm also more confident in my reasoning about conciliationism than my reasoning about free will. And yet my disagreement with Evelyn suggests that the former may rest on a

[23]Here's the analogy: As there appears to be nothing incoherent in the idea that we ought to act like utilitarians while believing nothing about the virtues of the utilitarian framework, there seems to be nothing incoherent in the idea that we ought to be conciliatory while suspending judgment on whether the norms prescribing conciliatory responses are correct—see (Littlejohn 2019, pp. 4–5).

[24]In an atypical case the agent is equally confident.

mistake. Clearly, my conclusion regarding conciliationism is either correct or not. If it is—in spite of the evidence to the contrary—I shouldn't trust my reasoning about free will. And if it is not, then the reasoning that I'm more confident in than my reasoning about free will has led me astray. Should I trust the reasoning I'm less confident in if the reasoning I was more confident in turned out to be mistaken? Perhaps, it's safer not to. Perhaps, in the end it is safer to suspend judgment on whether libertarian free will exists. Now, while one might find this train of thought overly cautious, it seems perfectly sensible and coherent! And this actually completes my response to the worry that conciliatory views can issue inconsistent (or incoherent) recommendations in scenarios involving disagreements about the correct way to disagree. It may be a problem for some conciliatory views, but not the one we have developed here.

There's one loose end left to tie up. At the outset of Chapter 4 we noted that the self-defeat problem breaks down into two sub-problems: the worry that conciliatory views can issue inconsistent directives and the worry that such views might self-defeat in the actual world. Now we have a response to the first worry. But what can we say in response to the second one?

The short answer is not that much. However, our formal analysis would seem to let us sharpen the statement of the worry, as well as to mitigate it somewhat, by showing that it's scope is not as wide as one might have thought. The worry itself boiled down to, you will recall, the claim that many—perhaps, even all—actual proponents of conciliationism find themselves in the sorts of circumstances where they

can't rationally hold onto their conciliatory views.[25] Now, if our formal analysis is on the right track, then one can't rationally believe that conciliationism is correct in case one's rational degree of confidence in one's reasoning about it is only as high as one's rational degree of confidence in there being a genuine disagreement about conciliationism—or in case $Seems(C) \leq Seems(Disagree(C))$. This means that an actual advocate of conciliationism is in trouble only in case she *should* be more confident in her disagreement with people like Kelly (2005, 2010), Titelbaum (2015), and Wedgwood (2010) than she *should* be in her own considerations about conciliationism. It seems intuitively plausible that there are some individuals whose relative degree of confidence in conciliationism and the disagreement about it are, respectively, higher and lower than they should be. These individuals, if there are any, are indeed being irrational in sticking to their conciliatory views. And it seems equally plausible that there are other individuals whose (relevant) degrees of confidence are close to what they should be and who are more confident in their considerations about conciliationism than the disagreement about it. These individuals are perfectly well justified in sticking to their conciliatory views. So, ultimately, our response to the second worry should, I think, run as follows: The answer to the question of whether some particular advocate of conciliationism is being irrational in holding onto her view will depend on the particular details of her epistemic situation, and it, thus, is an empirical matter.[26] Also, it would seem that, as long as the person in question can explain why she is more confident in the arguments that

[25]Decker (2014) thinks that this is the real problem that the Double Disagreement-like scenarios reveal.

[26]Cf. (Matheson 2015a, p. 149).

have led her to adopt conciliationism than in the arguments for the claim that her disagreement with the likes of Kelly is genuine, she is on safe grounds.[27]

## 6.5 Summary

Let's take a brief look back at this part of the dissertation. What we have done in it, in effect, is work out a view of conciliationism as a second-order defeasible reasoning policy saying, roughly, the following: If your best (first-order) reasoning suggests that $X$ and it's rational for you to think that you're a party to a genuine disagreement about whether $X$, you should not conclude that $X$ under normal circumstances. The phrases "it's rational for you to think" and "under normal circumstances" have precise content in the model. We also used the model to address one of the main challenges to conciliatory views: Given that there are disagreements about the correct way to disagree, conciliatory views would seem to self-defeat.

We noted that this challenge gives rise to two sub-problems. The first is that, through turning on themselves, conciliatory views would seem to issue inconsistent directives. The second is that actual advocates of conciliationism would seem to be irrational in holding onto their views. The bulk of the three chapters was concerned with working out a response to the first problem: Our first model conciliatory

[27]There's good reason for the advocates of conciliationism to be pleased with this result of our analysis. Some of have thought that all actual advocates of conciliationism are not holding their views rationally, which is the likely cause of the seemingly desperate suggestion that one can concede that one's view is not a view one can rationally believe without conceding that one's view is mistaken—see (Christensen 2009, p. 763) and (Littlejohn 2013, p. 175). (It's worth noting that in a later article Christensen admits that this is not a viable strategy—see (Christensen 2013, p. 82).) But our result is compatible with the claim that most actual advocates of concilaitionism are rational in sticking to their conciliatory views.

reasoner—based on default logic—appeared to corroborate the inconsistency worry, since it suggested that the correct (conciliatory) response to scenarios involving disagreements about the correct way to disagree is to conclude everything. Then, drawing on the tools of formal argumentation theory, we went on to formulate a model reasoner that generalized the original one. This more sophisticated reasoner suggested that the correct response to the unfortunate scenarios is to abandon the belief in conciliationism, all while responding to disagreements in the distinctively conciliatory way. Having pointed out that the apparent incoherence of this recommendation is due to the fact that the underlying scenarios were underdescribed, we proceeded to extend the reasoner in one final way, enabling it take into account (relative) degrees of confidence in the conclusions of the first-order, or domain-specific, reasoning. This put us in the position to represent the nuances of various scenarios involving disagreements about the correct way to disagree, and it allowed the reasoner to take these nuances into account when drawing conclusions. As we saw, while sometimes unorthodox, all of the reasoner's recommended responses were perfectly sensible, including the one that had appeared incoherent before. Thus, we have a response to the first problem. What's more, our analysis appears to mitigate the second problem too, suggesting that its scope is more narrow than initially thought.

# Part IV APPENDIX

**Observation 1** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a regular hedged context. Then there is a regular weighted context $c' = \langle \mathcal{W}', \mathcal{R}', \leq \rangle$ such that*

*(1) $\mathcal{W}' = \mathcal{W}$;*

*(2) for every rule $r' \in \mathcal{R}'$, there's a counterpart rule $r \in \mathcal{R}$;*

*(3) $X$ is a reason for $Y$ in $c$ if and only if $X$ is a reason for $Y$ in $c'$;*

*(4) $X$ is defeated as a reason for $Y$ by a consideration $Z$ in $c$ if and only if $X$ is defeated as a reason for $Y$ by $Z$ in $c'$; and*

*(5) $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.*

**Proof.**

Take an arbitrary regular hedged context $c = \langle \mathcal{W}, \mathcal{R} \rangle$. We will construct a (regular) weighted context from it. As a first step, we define an ordering on rules from $\mathcal{R}$, using the hedges of these rules, as follows. For any two rules $r, r' \in \mathcal{R}$, let

$$r \leq r' \qquad \text{if and only if} \qquad \neg Premise[r'] \in Hedge[r].$$

Now let $c'$ be the weighted context $\langle \mathcal{W}', \mathcal{R}', \leq \rangle$, where

- $\mathcal{W}' = \mathcal{W}$,
- $\mathcal{R}' = \{ \quad \frac{X}{Y} : \quad \frac{X : \mathcal{Z}}{\bigcirc Y} \in \mathcal{R} \}$, and
- $r \leq r'$ if and only if $counterpart_c(r) \leq counterpart_c(r')$.

**Claim 1**: $X$ is a reason for $Y$ in $c$ if and only if $X$ is a reason for $Y$ in $c'$.

Left-to-right: Suppose $X$ is a reason for $Y$ in $c$. Then there's a rule $r = \frac{X \,:\, \mathcal{Z}}{\bigcirc Y}$ in $\mathcal{R}$ such that $r \in Triggered(c)$. Since $r$ gets triggered in $c$, we have $\mathcal{W} \vdash X$. Given how $c'$ is constructed, we can be sure that there's a rule $r' \in \mathcal{R}'$ of the form $\frac{X}{Y}$. Since $\mathcal{W}' = \mathcal{W}$, the rule $r'$ must be triggered in $c'$, which suffices to conclude that $X$ is a reason for $Y$ in $c'$.

The other direction is similar.

**Claim 2**: $X$ is defeated as a reason for $Y$ by $Z$ in $c$ if and only if $X$ as a reason for $Y$ is defeated by $Z$ in $c'$.

Left-to-right: Suppose $X$ is defeated as a reason for $Y$ by $Z$ in $c$. This entails that there's a rule $r = \frac{X \,:\, \mathcal{Z}}{\bigcirc Y}$ in $\mathcal{R}$ such that $r \in Triggered(c)$. What's more, $\neg Z \in Hedge[r]$ and $\mathcal{W} \vdash Z$. In light of Claim 1, we know that there's a rule $r' = \frac{X}{Y}$ in $\mathcal{R}'$ that's triggered in $c'$, and, hence, that $X$ is a reason for $Y$ in $c'$. Now, Constraint (2) on regular hedged contexts tells us that there must be a rule $r^* \in \mathcal{R}$ such that $contrary_c(r, r^*)$ and $Premise[r^*] = Z$. By construction of $c'$, we can be sure that there's a rule $r'' \in \mathcal{R}'$ such that $counterpart_c(r'') = r^*$. What's more, we have $\mathcal{W}' \vdash Premise[r''] = Z$, $contrary_{c'}(r', r'')$, and $r' \leq r''$. (Why the latter? Well, it's entailed by the fact that $\neg Premise[r^*] \in Hedge[r]$.) And this is enough to conclude that $Z$ defeats $X$ as a reason for $Y$ in $c'$.

Right-to-left: Suppose $X$ is defeated as a reason for $Y$ by $Z$ in $c'$. This entails that there's a rule $r = \frac{X}{Y}$ in $Triggered(c')$ and a rule $r' = \frac{Z}{W}$ such that $r^* \in Triggered(c')$, $contrary_{c'}(r, r^*)$, and $r \leq r^*$. By construction, there are hedged rules $r'$ and $r'' \in \mathcal{R}$ such that $counterpart_c(r) = r'$, $counterpart_c(r^*) = r''$, and

$\neg Premise[r''] \in Hedge[r']$. But since $Premise[r''] = Z$ and $\mathcal{W} \vdash Z$, the formula $X$ gets defeated by $Z$ as a reason for $Y$ in $c$.

**Claim 3**: $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.

Left-to-Right: Suppose $\bigcirc X$ follows from $c$. This means that there's a rule $r \in \mathcal{R}$ such that $Conclusion[r] = \bigcirc X$ and $r \in Admissible(c)$, that is, $\mathcal{W} \vdash Premise[r]$ and there's no $\neg Z \in Hedge[r]$ such that $\mathcal{W} \vdash Z$. By the construction of $c'$, we know that there's a rule $r' \in \mathcal{R}'$ such that $r' = counterpart_{c'}(r)$. If $r' \in Binding(c')$, then we are done. So suppose the opposite. Given that $Premise[r] = Premise[r']$ and $\mathcal{W} \vdash Premise[r]$, it has to be the case that $r' \in Triggered(c')$. But given that it's not in $Binding(c')$, there must be another rule $r'' \in Triggered(c')$ such that $contrary_c(r', r'')$ and $r' \leq r''$. This means that $r'$ is defeated in the context of $c'$, and, in light of Claim 2, entails that $r$ must be defeated in the context of $c$, giving us a contradiction.

Right-to-left: Suppose that $\bigcirc X$ follows from $c'$. This means that there's a rule $r \in \mathcal{R}'$ such that $Conclusion[r] = X$ and $r \in Binding(c')$, which, in turn, means that $\mathcal{W}' \vdash Premise[r]$ and that there is no $r' \in \mathcal{R}'$ such that $r' \in Triggered(c')$ and $r \leq r'$. By construction of $c'$, we can be sure that there's a rule $r' \in \mathcal{R}$ such that $r' = counterpart_c(r)$. If we can show that $r' \in Admissible(c)$, then we are done. So let's suppose that it isn't. Given that $\mathcal{W}' \vdash Premise[r]$, it must hold that $\mathcal{W} \vdash Premise[r']$ and that $r' \in Triggered(c)$. So $r'$ is not admissible because there's a $Z$ such that $\neg Z \in Hedge[r]$ and $\mathcal{W} \vdash Z$. But, in light of Claim 2, this is enough to conclude that $r$ must be defeated in the context of $c$, giving us a contradiction.

QED

**Observation 2** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a regular weighted context. Then there is a regular hedged context $c' = \langle \mathcal{W}', \mathcal{R}' \rangle$ such that*

*(1) $\mathcal{W}' = \mathcal{W}$;*

*(2) for every rule $r' \in \mathcal{R}'$, there's a counterpart rule $r \in \mathcal{R}$;*

*(3) $X$ is a reason for $Y$ in $c$ if and only if $X$ is a reason for $Y$ in $c'$;*

*(4) $X$ is defeated as a reason for $Y$ by a consideration $Z$ in $c$ if and only if $X$ is defeated as a reason for $Y$ by $Z$ in $c'$; and*

*(5) $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.*

**Proof.**

Take an arbitrary regular weighted context $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$. We will construct a (regular) hedged context from it. Let $c' = \langle \mathcal{W}', \mathcal{R}' \rangle$, where

- $\mathcal{W}' = \mathcal{W}$, and
- $\mathcal{R}'$ is acquired from $(\mathcal{R}, \leq)$ by the following simple procedure.

For every rule $r \in \mathcal{R}$,

1. Let $\mathcal{R}_r = \{r' \in \mathcal{R} : r \leq r' \text{ and } contrary_{c^{\leq}}(r, r')\}$;
2. set $\mathcal{Z} = \{\neg X : \langle X, Y \rangle \in \mathcal{R}_r\}$;
3. and, finally, replace $r \in \mathcal{R}$ for the hedged rule

$$\frac{Premise[r] : \mathcal{Z}}{\bigcirc Conclusion[r]}.$$

**Claim 1**: $X$ is a reason for $Y$ in $c$ if and only if $X$ is a reason for $Y$ in $c'$.

Left-to-Right: Suppose $X$ is a reason for $Y$ in $c$. This means that there's a rule $r = \frac{X}{Y}$ in $\mathcal{R}$ and $r \in Triggered(c)$. By construction of $c'$, we can be sure that there's a rule $r' = \frac{X : \mathcal{Z}}{\bigcirc Y}$ in $\mathcal{R}'$ and that $\mathcal{W}' \vdash X$. Hence, $X$ is a reason for $Y$ in $c$ too.

The other direction is straightforward.

**Claim 2**: $X$ is defeated as a reason for $Y$ by a consideration $Z$ in $c$ if and only if $X$ is defeated as a reason for $Y$ by $Z$ in $c'$.

Left-to-Right: Suppose that $X$ is defeated as a reason for $Y$ by $Z$ in $c$. This entails that there's a rule $r = \frac{X}{Y}$ in $Triggered(c)$. Given how $c'$ is constructed, we can be sure that $r$ has a counterpart $r' = \frac{X : \mathcal{Z}}{\bigcirc Y}$ in $c'$ and that $r'$ is triggered in $c$. Since $X$ is defeated by $Z$, there must also be a rule $r^* = \frac{Z}{W}$ in $\mathcal{R}$ such that $r^* \in Triggered(c)$, $contrary_c(r, r^*)$, and $r \leq r^*$. By the construction of $c'$, we can be sure that $\neg Z \in Hedge[r']$ and that $\mathcal{W}' \vdash Z$. As a consequence, $X$ is defeated as a reason for $Y$ by $Z$ in $c'$.

The other direction is similarly straightforward.

**Claim 3**: $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.

Left-to-Right: Suppose that $\bigcirc X$ follows from $c$. This means that there's a rule $r \in \mathcal{R}$ such that $Conclusion[r] = X$ and $r \in Binding(c)$, with the latter fact implying that $\mathcal{W} \vdash Premise[r]$ and that there's no rule $r' \in Triggered(c)$ such that $contrary_c(r, r')$ and $r \leq r'$. Now, by the construction of $c'$, there's a rule $r' \in \mathcal{R}'$ with $r' = counterpart_{c'}(r)$. In case we can show that $r' \in Admissible(c')$, we are done. So suppose that it isn't. Since $\mathcal{W}' \vdash Premise[r']$, the rule $r'$ is not admissible because

of its hedge. So there's a $\neg Z \in Hedge[r']$ such that $\mathcal{W}' \vdash Z$. This, in turn, means that the rule $r'$ is defeated in $c'$, and, by Claim 2, we can be sure that $r$ too must be defeated in $c$. This gives us a contradiction.

Right-to-Left: Suppose $\bigcirc X$ follows from $c'$. This means that there's a rule $r \in \mathcal{R}'$ such that $Conclusion[r] = \bigcirc X$ and $r \in Admissible(c)$, that is, $\mathcal{W}' \vdash Premise[r]$ and, for no $\neg Z \in Hedge[r]$, do we have $\mathcal{W}' \vdash Z$. By construction, there has to be a rule $r' \in \mathcal{R}$ such that $counterpart_c(r) = r'$. If we can show that $r' \in Binding(c)$, then we are done. So suppose that it isn't. It's straightforward to see that $\mathcal{W} \vdash Premise[r']$. And, therefore, there must be a rule $r'' \in \mathcal{R}$ such that $r'' \in Triggered(c)$, $contrary_c(r', r'')$, and $r' \leq r''$. So $r'$ is actually defeated in $c$. By Claim 2, $r$ must be defeated in $c$ as well, contradicting the original assumption.

QED

**Observation 3** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a regular hedged context and $r$ a rule from $\mathcal{R}$ that's triggered in $c$. Then, if $r$ is defeated in $c$, it is rebutted.*

**Proof.**

Suppose $r$ is defeated in $c$. This means that there's some $Z$ such that $\neg Z \in Hedge[r]$ and $\mathcal{W} \vdash Z$. In light of the second constraint on regular hedged contexts, we know that $Hedge[r] \subseteq \{\neg Premise[r'] : r' \in \mathcal{R} \text{ and } contrary_c(r, r')\}$. So we can be sure that $Z = Premise[r']$ where $r'$ is a rule that's contrary to $r$. But, then, $r$ is rebutted by $Z$. QED

**Observation 4** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a hedged context subject to the following constraint: For any two rules $r$, $r' \in \mathcal{R}$ such that $contrary_c(r, r')$, either $\neg Premise[r'] \in$*

*$Hedge[r]$ or $\neg Premise[r] \in Hedged[r']$. Then there's a mixed context $c' = \langle \mathcal{W}', \mathcal{R}', \leq \rangle$ such that*

*(1) $\mathcal{W}' = \mathcal{W}$;*

*(2) for every rule $r' \in R'$, there's a counterpart rule $r \in R$;*

*(3) $X$ is a reason for $Y$ in $c$ if and only if $X$ is a reason for $Y$ in $c'$;*

*(4) $X$ is rebutted as a reason for $Y$ by a consideration $Z$ in $c$ if and only if $X$ is rebutted as a reason for $Y$ by $Z$ in $c'$;*

*(5) $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.*

**Proof.**

Take an arbitrary hedged context $c = \langle \mathcal{W}, \mathcal{R} \rangle$. We will construct a mixed context from it. As a first step, we define an ordering on rules from $\mathcal{R}$, using the hedges of these rules, as follows. For any two rules $r, r' \in \mathcal{R}$, let

$$r \preceq r' \qquad \text{if and only if} \qquad \neg Premise[r'] \in Hedge[r].$$

Now let $c$ be the mixed context $\langle \mathcal{W}', \mathcal{R}', \leq \rangle$, where

- $\mathcal{W}' = \mathcal{W}$,

- $\mathcal{R}'$ is acquired from $\mathcal{R}$ by the following procedure:

For every rule $r = \dfrac{X : \mathcal{Z}_{Old}}{\bigcirc Y}$ from $\mathcal{R}$,

1. Let $\mathcal{R}_r = \{r' \in \mathcal{R} : r \preceq r' \text{ and } contrary_{c^{\preceq}}(r, r')\}$;

2. set $\mathcal{Z} = \{\neg Premise[r'] : r' \in \mathcal{R}_r\}$;

3. and, finally, replace $r$ for the rule

$$\frac{X : \mathcal{Z}_{Old} \backslash \mathcal{Z}}{Y}.$$

- Finally, for any two $r$, $r' \in \mathcal{R}'$, set $r \leq r'$ if and only if $counterpart_c(r) \leq counterpart_c(r')$.

**Claim 1**: $X$ is a reason for $Y$ in $c$ if and only if $X$ is a reason for $Y$ in $c'$.

Left-to-right: Suppose that $X$ is a reason for $Y$ in $c$. This means that there's a rule $r = \frac{X : \mathcal{Z}}{\bigcirc Y}$ in $\mathcal{R}$ and that this rule is not undermined in $c$. This means that there's no $Z$ such that $\neg Z \in \mathcal{Z}$, $Z \neq Premise[r']$ for every rule $r'$ in $\mathcal{R}$, and $\mathcal{W} \vdash Z$. Given how $c'$ is constructed, there has to be a rule $r^* \in \mathcal{R}'$ of the form $\frac{X : \mathcal{Z}'}{Y}$. Since $\mathcal{W}' = \mathcal{W}$, this rule is triggered in $c'$. If we can show that it is not undermined $c'$, then we are done. So supposed it is undermined in $c'$. Then there has to be a some $Z$ such that $\neg Z \in \mathcal{Z}'$ and $\mathcal{W}' \vdash Z$. But given how $c'$ was constructed, $\neg Z$ can be an element of $\mathcal{Z}'$ only in case $Z \neq Premise[r']$ for every $r' \in \mathcal{R}$. Since $\mathcal{Z}' \subseteq \mathcal{Z}$ and $\mathcal{W}' = \mathcal{W}$, we also have $\neg Z \in \mathcal{Z}$ and $\mathcal{W} \vdash Z$. And this is a contradiction.

The other direction is similar.

**Claim 2**: $X$ is rebutted as a consideration for $Y$ by $Z$ in $c$ if and only if $X$ is rebutted as a consideration for $Y$ by $Z$ in $c'$.

Left-to-right: Suppose $X$ is rebutted as a reason for $Y$ by $Z$ in $c$. This entails that there's a rule $r = \frac{X : \mathcal{Z}}{\bigcirc Y}$ in $\mathcal{R}$ such that $r \in Triggered(c)$, and that this rule is not undermined in $c$. In light of Claim 1, we can be sure that $X$ is a reason for

$Y$ in $c$. So there's a rule $r' = \frac{X : \mathcal{Z}'}{Y}$ and there's no $Z$ such that $\neg Z \in \mathcal{Z}'$ and $\mathcal{W} \vdash Z$. Now, given that $X$ gets rebutted, there has to be rule $r^*$ in $\mathcal{R}$ such that $contrary_c(r, r^*)$, $\mathcal{W} \vdash Premise[r^*]$, and $\neg Premise[r^*] \in \mathcal{Z}$. By the construction of $c'$, there must be a corresponding rule $r'' \in \mathcal{R}'$ such that $counterpart_c(r'') = r^*$. From here, it's easy to see that $\mathcal{W} \vdash Premise[r'']$, $contrary_{c'}(r', r'')$, and $r' \leq r''$. And so the rule $r'$ is rebutted by $r''$ in $c'$. This, in turn, implies that $X$ is rebutted as a consideration for $Y$ by $Z$ in $c'$.

The other direction is, again, not much different.

**Claim 3**: $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.

Left-to-Right: Suppose $\bigcirc X$ follows from $c$. This means that there's a rule $r \in \mathcal{R}$ such that $Conclusion[r] = \bigcirc X$ and $r \in Admissible(c)$, that is, $\mathcal{W} \vdash Premise[r]$ and there's no $\neg Z \in Hedge[r]$ such that $\mathcal{W} \vdash Z$. Taking into account the way we constructed $c'$, we know that there's a rule $r' = \frac{Premise[r] : \mathcal{Z}'}{X}$ such that $r' = counterpart_{c'}(r)$. If $r' \in Optimal(c')$, then we are done. So suppose the opposite. Given that $Premise[r] = Premise[r']$ and $\mathcal{W} \vdash Premise[r]$, we can be sure that $r' \in Triggered(c')$. Since $r'$ is not among the rules in $Optimal(c')$, it must either be undermined or rebutted by another rule. Suppose that it is undermined, or that there's a $Z$ such that $\mathcal{W}' \vdash Z$ and $\neg Z \in \mathcal{Z}'$. Given that $\mathcal{Z}' \subseteq \mathcal{Z}$ and $\mathcal{W}' = \mathcal{W}$, we can conclude that $\mathcal{W} \vdash Z$ and $\neg Z \in \mathcal{Z}$. And this means that $r \notin Admissible(c)$. Suppose that $r'$ is rebutted by another rule. In light of Claim 2, we can conclude that $r$ must be rebutted in the context $c$. Either way we get a contradiction.

Right-to-left: Suppose that $\bigcirc X$ follows from $c'$. This means that there's a

rule $r = \frac{Y\ :\ \mathcal{Z}}{X}$ in $\mathcal{R}'$ such that $r \in Optimal(c')$, which means that $\mathcal{W}' \vdash Y$, that there's no $r' \in \mathcal{R}'$ such that $r' \in Triggered(c')$ and $r \leq r'$, and that there's no $Z$ such that $\neg Z \in \mathcal{Z}$ and $\mathcal{W} \vdash Z$. By construction of $c'$, we can be sure that there's a rule $r' = \frac{Y\ :\ \mathcal{Z}'}{\bigcirc X}$ in $\mathcal{R}$ such that $r' = counterpart_c(r)$. If we can show that $r' \in Admissible(c)$, then we are done. So let us suppose that it is not. Given that $\mathcal{W}' \vdash Y$, the $r'$ is triggered in $c$. So $r'$ is not admissible because there's a $Z$ such that $\neg Z \in \mathcal{Z}'$ and $\mathcal{W} \vdash Z$. There are two options now, either $\neg Z \in \mathcal{Z}$ or $\neg Z \notin \mathcal{Z}$. If the former, then the rule $r$ is undermined in $c'$, implying that $r$ is not optimal in $c'$. If $\neg Z \notin \mathcal{Z}$, then, by the construction of $c'$, there has to be a rule $r''$ in $\mathcal{R}'$ such that $contrary_{c'}(r, r'')$ and $r \leq r''$. But if that's the case, then $r$ cannot be among the optimal rules of $c'$. So we have a contradiction.

QED

**Observation 5** *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a mixed context subject to the following constraint: For any two rules $r$, $r' \in \mathcal{R}$ such that $contrary_c(r, r')$, either $r \leq r'$ or $r' \leq r$. Then there is a hedged context $c' = \langle \mathcal{W}', \mathcal{R}' \rangle$ such that*

*(1) $\mathcal{W}' = \mathcal{W}$;*

*(2) for every rule $r' \in R'$, there's a counterpart rule $r \in R$;*

*(3) $X$ is a reason for $Y$ in $c$ if and only if $X$ is a reason for $Y$ in $c'$;*

*(4) $X$ is rebutted as a reason for $Y$ by a consideration $Z$ in $c$ if and only if $X$ is rebutted as a reason for $Y$ by $Z$ in $c'$;*

*(5) $\bigcirc X$ follows from $c$ if and only if $\bigcirc X$ follows from $c'$.*

**Proof.** (Sketch)

Take an arbitrary mixed context $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$. We will construct a hedged context from it. Let $c' = \langle \mathcal{W}', \mathcal{R}' \rangle$, where

- $\mathcal{W}' = \mathcal{W}$, and

- $\mathcal{R}'$ is acquired from $(\mathcal{R}, \leq)$ by the following simple procedure.

For every rule $r = \dfrac{X : \mathcal{Z}_{Old}}{Y}$ from $\mathcal{R}$,

1. Let $\mathcal{R}_r = \{r' \in \mathcal{R} : r \leq r' \text{ and } contrary_{c^{\leq}}(r, r')\}$;

2. set $\mathcal{Z}_{New} = \{\neg Premise[r'] : r' \in \mathcal{R}_r\}$;

3. and, finally, replace $r \in \mathcal{R}$ for the hedged rule

$$\frac{X : \mathcal{Z}_{Old} \cup \mathcal{Z}_{New}}{\bigcirc Y} .$$

The proofs of clauses (2)–(4) run parallel to the proofs of Observations 1, 2 and 4.

QED

**Observation 6** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context in which no Reasonable-formulas occur. Then there exists a context $c' = \langle \mathcal{W}, \mathcal{R} \rangle$ where Reasonable-formulas do occur such that*

$$X \textit{ follows from } c \textit{ if and only if } X \textit{ follows from } c',$$

*for all $X$ in which the predicate Reasonable doesn't occur.*

**Proof.**

Take an arbitrary context $c = \langle \mathcal{W}, \mathcal{R} \rangle$. Now set $c'$ to be the context $\langle \mathcal{W}', \mathcal{R} \rangle$ where $\mathcal{W}' = \mathcal{W} \cup \{Reasonable(\mathfrak{r}) : \ r \in \mathcal{R}\}$.

Left-to-right: Take some arbitrary formula $X$ that follows from $c$, according to the original definition of consequence. Then we know that, for every proper scenario $\mathcal{S}$ based on $c$, we have $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$. Now zoom in on one such proper scenario $\mathcal{S}$. By the definition of the notion, for all $r \in \mathcal{S}$, we have $r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S})$, $r \notin Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S})$, and $r \notin Excluded_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. Now let's refocus on the new context $c'$. It's not difficult to see that $\mathcal{S}$ qualifies as a proper scenario based on $c'$. Since all the original information is present in $c'$, we can be sure that, for all $r \in \mathcal{S}$, we have $r \in Triggered_{\mathcal{W}',\mathcal{R}}(\mathcal{S})$, $r \notin Conflicted_{\mathcal{W}',\mathcal{R}}(\mathcal{S})$, and $r \notin Excluded_{\mathcal{W}',\mathcal{R}}(\mathcal{S})$. What's more, by the construction of $c'$, we know that, for every $r \in \mathcal{S}$, there's a formula of the form $Reasonable(\mathfrak{r})$ in the hard information of $c'$. Hence, for every $r \in \mathcal{S}$, we have $r \in Reasonable_{\mathcal{W}',\mathcal{R}}(\mathcal{S})$. So $\mathcal{S}$ is proper. The same applies to other proper scenarios based on $c$, and so $X$ follows from $c'$, according to the modified definition.

Right-to-left: Suppose that $X$ doesn't contain the predicate *Reasonable* and that $X$ follows from $c'$, according to the modified definition of consequence. Then for every proper scenario $\mathcal{S}$ based on $c'$, it holds that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$. Take an arbitrary $\mathcal{S}$. Then, by the definition of proper scenario, we know that, for all $\mathcal{R} \in \mathcal{S}$, $r \in Triggered_{\mathcal{W}',\mathcal{R}}(\mathcal{S})$, $r \notin Conflicted_{\mathcal{W}',\mathcal{R}}(\mathcal{S})$, and $r \notin Excluded_{\mathcal{W}',\mathcal{R}}(\mathcal{S})$. Since, by construction, $c'$ doesn't contain any information that doesn't have to do with the new predicate *Reasonable* and wouldn't be contained in $c$, it's easy to see that, for

all $r \in \mathcal{S}$, $r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S})$, $r \notin Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S})$, and $r \notin Excluded_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. So $\mathcal{S}$ qualifies as a proper scenario based on $c$, implying that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$. The same applies to the other proper scenarios of $c'$, and so $X$ follows from $c$, according to the original definition.

QED

**Observation 7** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $X$ an arbitrary formula. Then $X$ follows from $c$ in default logic, $c \mathrel{|\!\sim} X$, if and only if $X$ follows from $c$ according to stability semantics, $c \mathrel{|\!\sim}_s X$.*

**Proof.**

Left-to-right: Suppose that $c \mathrel{|\!\sim} X$. The, for every proper scenario $\mathcal{S}$ based on $c$, we have $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$. Let's zoom in one such $\mathcal{S}$. Since $\mathcal{S} \subseteq Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S})$ and $\mathcal{S} \subseteq Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S})$, the set $\mathcal{S}$ is an element of $Arguments(c)$. Now we will show that that $\mathcal{S}$ defeats every argument $\mathcal{S}'$ in $Arguments(c)$ such that $\mathcal{S}' \notin \mathcal{S}$. So consider an arbitrary $\mathcal{S}'$ of this sort. Zoom in on $\mathcal{S}'' = \mathcal{S} \cap \mathcal{S}'$. Now take some rule $r$ from $\mathcal{S}'$ such that $r \in Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S}'')$, $r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S}'')$. Such an $r$ has to exist because $\mathcal{S}' \in Argument(c)$ and $\mathcal{S}'' \subset \mathcal{S}'$. In light of the fact that $r \in Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S}'')$ and $r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S}'')$, it has to be the case that $r \in Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S})$, $r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. So, given that $\mathcal{S}$ is a proper scenario, it must be the case that either $r \in Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S})$ or $r \in Excluded_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. So either $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$ or $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(\mathfrak{r})$. But in either case we get $\mathcal{S} \rightsquigarrow \mathcal{S}'$. Set $\Gamma = \{\mathcal{S}' \in Arguments(c) : \mathcal{S}' \subseteq \mathcal{S}\}$. Since $\mathcal{S}$ defeats every $\mathcal{S}'$ in $Arguments(c)$, the set of arguments $\Gamma$ defeats every argument it

doesn't contain. And given that $\mathcal{S}$ is a proper scenario, $\Gamma$ has to be consistent. So $\Gamma$ is a stable extension of $\mathcal{F}(c)$. What's more, $X$ follows from $\Gamma$, as it contains $\mathcal{S}$ and $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$. Notice that we can run the same argument for every other proper scenario based on $c$. Consequently, $c \mathrel{\sim\!\!\!\vdash_s} X$.

Right-to-left: Suppose that $c \mathrel{\sim\!\!\!\vdash_s} X$. This means that, for every stable extension $\Gamma$ of $\mathcal{F}(c)$, it holds that $\Gamma$ contains some argument $\mathcal{S}$ such that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$. Let's now focus on one such stable extension $\Gamma$.

The first step is to show that this $\Gamma$ has a maximal element, that is, an argument $\mathcal{S}$ such that, for all $\mathcal{S}' \in \Gamma$, we have $\mathcal{S}' \subseteq \mathcal{S}$. To show that this holds, we use a proof by contradiction. Suppose that there's no single maximal element in $\Gamma$. Now take some $\mathcal{S} \in \Gamma$ such that there's no $\mathcal{S}' \in \Gamma$ with $\mathcal{S} \subset \mathcal{S}'$. Consider an arbitrary rule $r$ from $\mathcal{R}$ such that $r \notin \mathcal{S}$, but $r \in Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S})$ and $r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. Since $\mathcal{S}$ is maximal, $\mathcal{S} \cup \{r\} \notin \Gamma$. And given that $\Gamma$ is stable, it must hold that $\Gamma \rightsquigarrow \mathcal{S} \cup \{r\}$. So there has to be some argument $\mathcal{S}' \in \Gamma$ such that $\mathcal{S}' \rightsquigarrow \mathcal{S} \cup \{r\}$. The expression $\mathcal{S}' \rightsquigarrow \mathcal{S} \cup \{r\}$ means that there has to be some rule $r'$ in $\mathcal{S} \cup \{r\}$ such that $\mathcal{W} \cup Conclusion[\mathcal{S}'] \vdash \neg Conclusion[r']$ or $\mathcal{W} \cup Conclusion[\mathcal{S}'] \vdash \neg Out(\mathfrak{r}')$. However, if the rule $r'$ in question is anything but $r$ itself, then we would also have $\mathcal{S}' \rightsquigarrow \mathcal{S}$, making $\Gamma$ inconsistent. So the argument $\mathcal{S}'$ is such that either $\mathcal{W} \cup Conclusion[\mathcal{S}'] \vdash \neg Conclusion[r]$ or $\mathcal{W} \cup Conclusion[\mathcal{S}'] \vdash \neg Out(\mathfrak{r})$. Notice that has to be such an argument $\mathcal{S}'$ for every $r$ with $r \notin \mathcal{S}$, but $r \in Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S})$ and $r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S})$—if there are any such rules at all.

Now let's zoom in on a different set $\mathcal{S}^\dagger \in \Gamma$ such that there's no $\mathcal{S}' \in \Gamma$ with $\mathcal{S}^\dagger \subset \mathcal{S}$. So $\mathcal{S}^\dagger \neq \mathcal{S}$. Consider $\mathcal{S} \cap \mathcal{S}^\dagger$. Take the rule $r^\dagger$ such that $r^\dagger \in \mathcal{S}^\dagger$,

$r^\dagger \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S} \cap \mathcal{S}^\dagger)$, and $r^\dagger \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S} \cap \mathcal{S}^\dagger)$. Since $\mathcal{S}$ and $\mathcal{S}^\dagger$ are both in $Arguments(c)$ and $\mathcal{S}^\dagger \notin \mathcal{S}$, such a rule $r^\dagger$ must exist. But given the proof in the previous paragraph, we can be sure that has to be an argument $\mathcal{S}' \in \Gamma$ such that either $\mathcal{W} \cup Conclusion[\mathcal{S}'] \vdash \neg Conclusion[r^\dagger]$ or $\mathcal{W} \cup Conclusion[\mathcal{S}'] \vdash Out(\mathfrak{r}^\dagger)$. Since $r^\dagger \in \mathcal{S}^\dagger$, we have $\mathcal{S} \rightsquigarrow \mathcal{S}^\dagger$ which entails, contrary to our assumption, that $\Gamma$ is inconsistent. So $\Gamma$ must have a maximal element after all.

It's not difficult to see that the maximal element of $\Gamma$, call it, $\mathcal{S}$, is such that, for all $\mathcal{S}' \in Argument(c)$ with $\mathcal{S}' \notin \mathcal{S}$, it holds that $\mathcal{S} \rightsquigarrow \mathcal{S}'$. Consider some $\mathcal{S}'$ that fits the description. Given that $\Gamma$ is stable, we know that $\Gamma \rightsquigarrow \mathcal{S}'$. So there's some argument $\mathcal{S}'' \in \Gamma$ such that $\mathcal{S}'' \rightsquigarrow \mathcal{S}'$, meaning that $\mathcal{W} \cup Conclusion[\mathcal{S}''] \vdash \neg Conclusion[r]$ or $\mathcal{W} \cup Conclusion[\mathcal{S}''] \vdash Out(\mathfrak{r})$ for some rule $r \in \mathcal{S}'$. But since $\mathcal{S}$ is maximal, it must be the case that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$ or $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(\mathfrak{r})$ for some rule $r \in \mathcal{S}'$. Consequently, $\mathcal{S} \rightsquigarrow \mathcal{S}'$. Another thing that should be clear is that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$: If there's an argument in $\Gamma$ that lets us conclude $X$, then $X$ follows from the maximal element too.

The final step is to show that the maximal element $\mathcal{S}$ of $\Gamma$ is a proper scenario based on $c$. What we need to establish, then, is that

$$\begin{aligned} \mathcal{S} \;=\; \{r \in \mathcal{R} :\; & r \in Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S}), \\ & r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S}), \\ & r \notin Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S}), \\ & r \notin Excluded_{\mathcal{W},\mathcal{R}}(\mathcal{S})\}. \end{aligned}$$

$\subseteq$ : Take an arbitrary $r$ from $\mathcal{S}$. Since $\mathcal{S}$ is in $Argument(c)$, we know that $r \in Reasonsable_{\mathcal{W},\mathcal{R}}(\mathcal{S})$ and $r \in Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. Now supposed there was a rule $r \in Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. In that case, $\mathcal{S}$ would self-defeat, and $\Gamma$ couldn't be a stable extension. So $r \notin Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. Analogous considerations apply to the possibility of $r$ being excluded in the context of $\mathcal{S}$.

$\supseteq$ : Take an arbitrary rule $r$ such that $r$ is an element of $Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S})$ and $Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S})$ and $r$ is not an element of either $Conflicted_{\mathcal{W},\mathcal{R}}(\mathcal{S})$, or $Excluded_{\mathcal{W},\mathcal{R}}(\mathcal{S})$. Now suppose, toward a contradiction, that $r \notin \mathcal{S}$. Let $\mathcal{S}' = \mathcal{S} \cup \{r\}$. Since $\mathcal{S}' \subseteq Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S}')$ and $\mathcal{S}' \subseteq Reasonable_{\mathcal{W},\mathcal{R}}(\mathcal{S}')$, $\mathcal{S}'$ must be in $Argument(c)$. What's more, we do not have $\mathcal{S} \rightsquigarrow \mathcal{S}$, and so $\Gamma \not\rightsquigarrow \mathcal{S}'$. This is enough to conclude that $\Gamma$ is not a stable extension after all.

This shows that $\mathcal{S}$ is proper. Since we can run the same argument for every other stable extension, we know that $c \vdash X$.

QED

**Observation 8** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{F}(c) = \langle \mathcal{A}, \rightsquigarrow \rangle$ an argument framework constructed from it. If $\mathcal{F}(c)$ does not contain either odd cycles of defeat or infinite chains of defeat, then $c \vdash_s X$ if and only if $c \vdash_p X$.*

**Proof.**

We will show that an argument set $\Gamma$ is a stable extension of $\mathcal{F}(c)$ if and only if it is a preferred extensions of $\mathcal{F}(c)$. The result follows immediately.

Left-to-right (Dung 1995): This direction holds independently of the assumption. Let $\Gamma$ be a stable extension of $\mathcal{F}(c)$. So, for all $\mathcal{S} \in \mathcal{A} \backslash \Gamma$, $\Gamma \rightsquigarrow \mathcal{S}$. It's easy to

see that $\Gamma$ is complete: Consider an argument $\mathcal{S}$ such that $\Gamma$ defends $\mathcal{S}$. If $\mathcal{S} \in \Gamma$, we're done. So suppose $\mathcal{S} \notin \Gamma$. Since $\Gamma$ is stable, we have it that $\Gamma \rightsquigarrow \mathcal{S}$. Thus, there's an argument $\mathcal{S}' \in \Gamma$ such that $\mathcal{S}' \rightsquigarrow \mathcal{S}$. Given that $\Gamma$ defends $\mathcal{S}$, there has to be an $\mathcal{S}^{\dagger}$ in $\Gamma$ such that $\mathcal{S}^{\dagger} \rightsquigarrow \mathcal{S}'$. But this would mean that $\Gamma$ is not conflict-free, which contradicts it being stable. Now let's verify that $\Gamma$ is not only a complete extension, but also a maximal complete extensions: Suppose that it wasn't. There would be another complete extension $\Gamma'$ such that $\Gamma \subset \Gamma'$. Let $\mathcal{S}$ be an argument such that $\mathcal{S} \notin \Gamma$ and $\mathcal{S} \in \Gamma'$. Since $\Gamma$ is stable, $\Gamma \rightsquigarrow \mathcal{S}$, and so $\Gamma' \rightsquigarrow \mathcal{S}$. Then, however, $\Gamma'$ is not conflict-free, which contradicts it being complete.

Right-to-left: Suppose that $\Gamma$ is a preferred, but not a stable extension of $\mathcal{F}(c)$. So $\Gamma$ is a maximal complete extension, and yet there is some argument $\mathcal{S}_1 \in \mathcal{A}$ such that $\mathcal{S}_1 \notin \Gamma$ and $\Gamma \not\rightsquigarrow \mathcal{S}_1$. Since $\Gamma$ is complete and $\mathcal{S}_1$ is not in $\Gamma$, there has to be an argument $\mathcal{S}_2$ such that $\mathcal{S}_2 \rightsquigarrow \mathcal{S}_1$ and $\Gamma \not\rightsquigarrow \mathcal{S}_2$. Either $\mathcal{S}_2 \in \Gamma$ or $\mathcal{S}_2 \notin \Gamma$. If the former, $\Gamma \rightsquigarrow \mathcal{S}_1$. So $\mathcal{S}_2 \notin \Gamma$. Since $\Gamma$ is complete and $\mathcal{S}_2$ is not in $\Gamma$, there has to be an argument $\mathcal{S}_3$ such that $\mathcal{S}_3 \rightsquigarrow \mathcal{S}_2$ and $\Gamma \not\rightsquigarrow \mathcal{S}_3$. This means that there's the following chain of defeat in $\mathcal{F}(c)$: $\mathcal{S}_3 \rightsquigarrow \mathcal{S}_2 \rightsquigarrow \mathcal{S}_1$. Either $\mathcal{S}_3 \in \Gamma$, or $\mathcal{S}_3 \notin \Gamma$. If $\mathcal{S}_3$ is in $\Gamma$, then $\Gamma \rightsquigarrow \mathcal{S}_2$, contradicting a fact established before. So $\mathcal{S}_3 \notin \Gamma$. Since $\Gamma$ is complete and $\mathcal{S}_3 \notin \Gamma$, there has to be an argument $\mathcal{S}_4$ such that $\mathcal{S}_4 \rightsquigarrow \mathcal{S}_3$ and $\Gamma \not\rightsquigarrow \mathcal{S}_4$. And we can apply the same line of reasoning to $\mathcal{S}_4$ and further, but it has to stop eventually, given that, by assumption, there are no infinitely ascending chains of defeat. So we will end up with the following possibly very long, but finite chain:

$$\mathcal{S}_n \rightsquigarrow \mathcal{S}_{n-1} \rightsquigarrow \cdots \rightsquigarrow \mathcal{S}_3 \rightsquigarrow \mathcal{S}_2 \rightsquigarrow \mathcal{S}_1.$$

and we will have established on the way that, for all $i$ with $1 \leq i < n$, $S_i \notin \Gamma$. But given that $\Gamma$ is preferred, $\mathcal{S}_n \in \Gamma$.

QED

**Observation 9** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ and $\mathcal{S}'$ arguments from the argument framework $\mathcal{F}(c)$ constructed from it. Then $\mathcal{S}$ defeats $\mathcal{S}'$, according to Definition 5.2, $\mathcal{S} \rightsquigarrow \mathcal{S}'$, if and only if $\mathcal{S}$ defeats $\mathcal{S}'$, according to Definition 5.12, $\mathcal{S} \rightsquigarrow_a \mathcal{S}'$.*

**Proof.**

Right-to-left: Suppose that $\mathcal{S} \rightsquigarrow_a \mathcal{S}'$. This implies that there are arguments $\mathcal{S}^\dagger$ and $\mathcal{S}^\ddagger$ in the set $Arguments(c)$ such that $\mathcal{S}^\dagger \subseteq \mathcal{S}$, $\mathcal{S}^\ddagger \subseteq \mathcal{S}'$, and $\mathcal{S}^\dagger \rightsquigarrow_b \mathcal{S}^\ddagger$. From here, either $\mathcal{W} \cup Conclusion[\mathcal{S}^\dagger] \vdash \neg Conclusion[r]$ or $\mathcal{W} \cup Conclusion[\mathcal{S}^\dagger] \vdash Out(\mathfrak{r})$ for some rule $r$ from $\mathcal{S}^\ddagger$. Since $\mathcal{S}^\dagger \subseteq \mathcal{S}$ and $\mathcal{S}^\ddagger \subseteq \mathcal{S}'$, it follows that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$ or $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(\mathfrak{r})$ for some rule $r$ from $\mathcal{S}'$. And this is enough to conclude that $\mathcal{S} \rightsquigarrow \mathcal{S}'$.

Left-to-right: Suppose that $\mathcal{S} \rightsquigarrow \mathcal{S}'$. This means that there is a rule $r \in \mathcal{S}'$ such that either $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash \neg Conclusion[r]$, or $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(\mathfrak{r})$. Without loss of generality, suppose that $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(\mathfrak{r})$. Now take the (set-theoretically) smallest argument $\mathcal{S}^\dagger$ in $Arguments(c)$ such that $\mathcal{W} \cup Conclusion[\mathcal{S}^\dagger] \vdash Out(\mathfrak{r})$ and $\mathcal{S}^\dagger \subseteq \mathcal{S}$. Since $\mathcal{S}$ is in $Arguments(c)$, we know that $\mathcal{S}^\dagger$ exists. It's easy to see that $\mathcal{S}^\dagger$ is in the set $Minimal_{\mathcal{F}(c)}(Out(\mathfrak{r}))$: If not, then there must be another set $\mathcal{S}^\ddagger \subset \mathcal{S}^\dagger$ in $Arguments(c)$ such that $\mathcal{W} \cup Conclusion[\mathcal{S}^\ddagger] \vdash Out(\mathfrak{r})$. In that case, however, we'd also have $\mathcal{S}^\ddagger \subset \mathcal{S}$, contradicting our assumption that $\mathcal{S}^\dagger$ is the smallest arguments that's also a subset of $\mathcal{S}$ that entails $Out(\mathfrak{r})$

with $\mathcal{W}$. Given our definition of basic defeat, $\mathcal{S}^\dagger \rightsquigarrow_b \mathcal{S}''$, for any $\mathcal{S}''$ such that $\mathcal{S}''$ is in $Minimal_{\mathcal{F}(c)}(r)$. Let $\mathcal{S}^\ddagger$ be the (set-theoretically) smallest argument from $Arguments(c)$ with both $r \in \mathcal{S}^\ddagger$ and $\mathcal{S}^\ddagger \subseteq \mathcal{S}'$. Since $\mathcal{S}'$ is in $Arguments(c)$, we can be sure that $\mathcal{S}^\ddagger$ exists. It's, again, easy to see that $\mathcal{S}^\ddagger$ is among the arguments in $Minimal_{\mathcal{F}(c)}(r)$, from which it follows that $\mathcal{S}^\dagger \rightsquigarrow_b \mathcal{S}^\ddagger$. Finally, given that $\mathcal{S}^\dagger \subseteq \mathcal{S}$ and $\mathcal{S}^\ddagger \subseteq \mathcal{S}'$, we also have $\mathcal{S} \rightsquigarrow_a \mathcal{S}'$. QED

**Observation 10** *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be an ordinary context and $c' = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be the same context with a connected preorder $\leq$ assigning all the rules $r$ in $\mathcal{R}$ the same weight—so, for all $r$, $r' \in \mathcal{R}$, $r \sim r'$. Then $\mathcal{F}(c) = \mathcal{F}(c')$.*

**Proof.**

The sets of arguments of $\mathcal{F}(c)$ and $\mathcal{F}(c')$ are clearly the same. So it remains to show that the defeat relations among the arguments in them coincide.

Left-to-right: Take two arbitrary arguments $\mathcal{S}$ and $\mathcal{S}'$ from $\mathcal{F}(c)$ with $\mathcal{S} \rightsquigarrow \mathcal{S}'$. In light of Observation 9, we know that $\mathcal{S} \rightsquigarrow_a \mathcal{S}'$. So the set $Arguments(c)$ must contain two arguments $\mathcal{S}^\dagger$ and $\mathcal{S}^\ddagger$ such that $\mathcal{S}^\dagger \subseteq \mathcal{S}$, $\mathcal{S}^\ddagger \subseteq \mathcal{S}'$, and $\mathcal{S}^\dagger \rightsquigarrow_b \mathcal{S}^\ddagger$. This, in turn, means that $\mathcal{S}^\ddagger$ is an element of $Minimal_{\mathcal{F}(c')}(r)$, as well as that either (i) $\mathcal{W} \cup Conclusion[\mathcal{S}^\dagger] \vdash \neg Conclusion[r]$ and $\mathcal{S}^\dagger$ is in $Minimal_{\mathcal{F}(c')}(\neg Conclusion[r])$, or (ii) $\mathcal{W} \cup Conclusion[\mathcal{S}^\dagger] \vdash Out(\mathfrak{r})$ and $\mathcal{S}^\dagger$ is in $Minimal_{\mathcal{F}(c')}(Out(\mathfrak{r}))$. Since in $c'$ all rules are assigned the same weight, for all $r \in \mathcal{S}^\ddagger$ and all $r' \in \mathcal{S}^\dagger$, $r \leq r'$. And given that $\mathcal{S}^\ddagger$ is not empty, there is some rule $r \in \mathcal{S}^\ddagger$ such that $r \leq r'$ for every rule $r' \in \mathcal{S}^\dagger$. This is enough to conclude that $\mathcal{S}^\ddagger \preceq \mathcal{S}^\dagger$. From here, we have $\mathcal{S}^\dagger \rightsquigarrow_b^w \mathcal{S}^\ddagger$ and, after another step $\mathcal{S} \rightsquigarrow^w \mathcal{S}'$.

The right-to-left direction is analogous.

QED

# Bibliography

Barberà, S., Bossert, W., & Pattaniak, P. (2004). Ranking sets of objects. In S. Barberà, P. Hammond, & C. Seidl (Eds.), *Handbook of Utility Theory, Volume 2* (pp. 893–977). Springer Science+Business Media.

Barringer, H., Gabbay, D., & Woods, J. (2012). Temporal, numerical and meta-level dynamics in argumentation networks. *Argument and Computation*, 3(2–3), 143–202.

Bogardus, T. (2009). A vindication of the Equal-Weight View. *Episteme*, 6(3), 324–35.

Boghossian, P. A. (2017). Epistemic rules. *Journal of Philosophy*, 105(9), 472–500.

Bradley, D. (2019). Are there indefeasible epistemic rules? *Philosopher's Imprint*, 19(3), 1–19.

Brass, S. (1991). Deduction with supernormal defaults. In G. Brewka, K. Jantke, & P. Schmitt (Eds.), *Nonmonotonic and Inductive Logics. Lecture Notes in Computer Science, Volume 659* (pp. 153–74).: Springer.

Broome, J. (1999). Normative requirements. *Ratio*, 12, 398–419.

Broome, J. (2007). Wide or narrow scope? *Mind*, 116(462), 359–70.

Broome, J. (2013). *Rationality through Reasoning*. Wiley Blackwell Publishing.

Brown, C. (2014). The composition of reasons. *Synthese*, 191, 779–800. DOI 10.1007/s 11229-013-0299-8.

Chisholm, R. (1964). The ethics of requirement. *American Philosophical Quarterly*, 1(2), 147–53.

Chisholm, R. (1980). A version of foundationalism. *Midwest Studies in Philosophy*, 5(1), 543–64.

Christensen, D. (2007a). Does Murphy's Law apply in epistemology? Self-doubt and rational ideals. *Oxford Studies in Epistemology*, 2, 3–31.

Christensen, D. (2007b). Epistemology of disagreement: The good news. *Philosophical Review*, 116, 187–217.

Christensen, D. (2009). Disagreement as evidence: The epistemology of controversy. *Philosophy Compass*, 4, 756–67.

Christensen, D. (2010a). Higher-order evidence. *Philosophy and Phenomenological Research*, 81(1), 185–215.

Christensen, D. (2010b). Rational reflection. *Philosophical Perspectives*, 24, 121–140.

Christensen, D. (2011). Disagreement, question-begging and epistemic self-criticism. *Philosophers' Imprint*, 11(6), 1–22.

Christensen, D. (2013). Epistemic modesty defended. In D. Christensen & J. Lackey (Eds.), *The Epistemology of Disagreement: New Essays* (pp. 77–97). Oxford University Press.

Christensen, D. (2016). Conciliation, uniqueness and rational toxicity. *Noûs*, 50(3), 584–603.

Coates, A. (2012). Rational epistemic akrasia. *American Philosophical Quarterly*, 49(2), 113–24.

Conee, E. & Feldman, R. (2004). *Evidentialism: Essays in Epistemology*. Oxford University Press.

Conee, E. & Feldman, R. (2008). Evidence. In Q. Smith (Ed.), *Epistemology: New Essays* (pp. 83–104). Oxford University Press.

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

Dancy, J. (2017). Moral particularism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition.

Decker, J. (2014). Conciliation and self-incrimination. *Erkenntnis*, 79, 1099–134.

Delgrande, J. & Schaub, T. (2004). Reasoning with sets of defaults in default logic. *Computational Intelligence*, 20, 56–88.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2), 321–57.

Dunne, P., Hunter, A., McBurney, P., Parsons, S., & Wooldridge, M. (2011). Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2), 457–86.

Elga, A. (2007). Reflection and disagreement. *Noûs*, 41(3), 478–502.

Elga, A. (2010). How to disagree about how to disagree. In R. Feldman & T. Warfield (Eds.), *Disagreement* (pp. 175–86). Oxford University Press.

Feldman, R. (2005). Respecting the evidence. *Philosophical Perspectives*, 19(1), 95–119.

Feldman, R. (2006). Epistemological puzzles about disagreement. In S. Hetherington (Ed.), *Epistemology Futures* (pp. 216–36). Oxford University Press.

Feldman, R. (2009). Evidentialism, higher-order evidence, and disagreement. *Episteme*, 6, 294–312.

Fogal, D. (f). Rational requirements and the primacy of pressure. *Mind*, forthcoming.

Foot, P. (1983). Moral realism and moral dilemma. *Journal of Philosophy*, 80, 379–98.

Gabbay, D. (2012). An equational approach to argumentation networks. *Argument and Computation*, 3(2–3), 87–142.

Gelfert, A. (2011). Who is an epistemic peer? *Logos and Episteme*, 2(4), 507–14.

Goble, L. (2009). Normative conflicts and the logic of 'ought'. *Noûs*, 43:3, 450–89.

Gómez Lucero, M. J., Chesñevar, C. I., & Simari, G. R. (2009). Modelling argument accrual in possibilistic defeasible logic programming. In C. Sossai & G. Chemello (Eds.), *Symbolic and quantitative approaches to reasoning with uncertainty. ECSQARU. Lecture Notes in Computer Science, Volume 5590* (pp. 131–43).: Springer.

Gómez Lucero, M. J., Chesñevar, C. I., & Simari, G. R. (2013). Modelling argument accrual with possibilistic uncertainty in a logic programming setting. *Information Sciences*, 228, 1–25.

Gowans, C. (1987). Introduction: The debate on moral dilemmas. In C. Gowans (Ed.), *Moral Dilemmas* (pp. 3–33). Oxford University Press.

Grossi, D. & Modgil, S. (2015). On the graded acceptability of arguments. In *Proceedings of the 24th Joint Conference on Artificial Intelligence, IJCAI* (pp. 868–74).: AAAI Press.

Holton, R. (2002). Principles and particularisms. In *Proceedings of the Aristotelian Society, Suppl. Volume 76* (pp. 191–210).

Horowitz, S. (2014). Epistemic akrasia. *Noûs*, 48(4), 718–4.

Horty, J. F. (1994). Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23, 35–65.

Horty, J. F. (2003). Reasoning with moral conflicts. *Noûs*, 37, 557–605.

Horty, J. F. (2012). *Reasons as Defaults*. Oxford University Press.

Huemer, M. (2000). Direct realism and the brain-in-a-vat argument. *Philosophy and Phenomenological Research*, 88(2), 397–413.

Hughes, N. (2017). Dilemmic epistemology. *Synthese*. https://doi.org/10.1007/s11229-017-1639-x.

Kelly, T. (2005). The epistemic significance of disagreement. *Oxford Studies in Epistemology*, 1, 179–92.

Kelly, T. (2010). Peer disagreement and higher-order evidence. In R. Feldman & T. Warfield (Eds.), *Disagreement* (pp. 111–74). Oxford University Press.

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

Kolodny, N. (2005). Why be rational? *Mind*, 114, 509–60.

Lackey, J. (2010a). A justificationists view of disagreement's epistemic significance. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Social Epistemology* (pp. 298–325). Oxford University Press.

Lackey, J. (2010b). What should we do when we disagree? *Oxford Studies in Epistemology*, 3, 274–93.

Lance, M. & Little, M. (2007). Where the laws are. *Oxford Studies in Metaethics*, 2, 149–71.

Lasonen-Aarnio, M. (2013). Disagreement and evidential attenuation. *Noûs*, 47, 767–94.

Lasonen-Aarnio, M. (2014). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, 88(2), 314–45.

Lasonen-Aarnio, M. (2020). Enkrasia or evidentialism? Learning to love mismatch. *Philosophical Studies*, 177, 597–632.

Leonard, N. (2020). Epistemic dilemmas and rational indeterminacy. *Philosophical Studies*, 177, 573–96.

Littlejohn, C. (2013). Disagreement and defeat. In D. Machuca (Ed.), *Disagreement and Skepticism* (pp. 169–92). Routledge.

Littlejohn, C. (2018). Stop making sense? On a puzzle about rationality. *Philosophy and Phenomenological Research*, 96(2).

Littlejohn, C. (2019). Should we be dogmatically conciliatory? *Philosophical Studies*. https://doi.org/10.1007/s11098-019-01258-4.

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

Makinson, D. (2005). *Bridges from Classical to Nonmonotonic Logic*. King's College Publications.

Makinson, D. & van der Torre, L. (2000). Input/output logics. *Journal of Philosophical Logic*, 29, 383–408.

Makinson, D. & van der Torre, L. (2001). Constraints for input/output logics. *Journal of Philosophical Logic*, 30, 155–85.

Matheson, J. (2015a). Are conciliatory views of disagreement self-defeating? *Social Epistemology*, 29(2), 145–59.

Matheson, J. (2015b). Epistemic norms and self-defeat: A reply to littlejohn. *Social Epistemology Review and Reply Collective*, 4(2), 26–32. http://wp.me/p1Bfg0-1Uo.

Matheson, J. (2015c). *The Epistemic Significance of Disagreement*. Pelgrave Macmillan.

Matheson, J. (2018). Disagreement and epistemic peers. In *Oxford Handbooks Online*. DOI: 10.1093/oxfordhb/97801999353.

McCain, K. (2014). *Evidentialism and Epistemic Justification*. New York: Routledge.

McKeever, S. & Ridge, M. (2006). *Principled Ethics: Generalism as a Regulative Ideal*. Oxford University Press.

McNamara, P. (2019). Deontic logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition.

Modgil, S. & Prakken, H. (2013). A general account of argumentation with preferences. *Artificial Intelligence*, 195, 361–97.

Mulligan, T. (2015). Disagreement, peerhood, and three paradoxes of conciliationism. *Synthese*, 192, 67–78.

Nair, S. (2016). How do reasons accrue? In E. Lord & B. Maguire (Eds.), *Weighing Reasons* (pp. 56–73). Oxford University Press.

Parent, X. (2011). Moral particularism in the light of deontic logic. *Artificial Intelligence and Law*, 19, 75–98.

Pittard, J. (2015). Resolute conciliationism. *The Philosophical Quarterly*, 65(260), 442–63.

Pollock, J. (1974). *Knowledge and Justification*. Princeton: Princeton University Press.

Pollock, J. (1994). Justification and defeat. *Artificial Intelligence*, 67, 377–407.

Pollock, J. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA: MIT Press.

Pollock, J. (2001). Defeasible reasoning with variable degrees of justification. *Artificial Intelligence*, 133, 233–82.

Pollock, J. (2008). Irrationality and cognition. In Q. Smith (Ed.), *Epistemology: New Essays* (pp. 249–75). Oxford University Press.

Pollock, J. (2010). Defeasible reasoning and degrees of justification. *Argument and Computation*, 1(1), 7–22.

Pollock, J. & Cruz, J. (1999). *Contemporary Theories of Knowledge*. Rowman & Littlefield Publishers.

Prakken, H. (2005). A study of accrual of arguments, with applications to evidential reasoning. In *Proceedings of the 10th International Conference on Artificial Intelligence and the Law* (pp. 85–94).

Prakken, H. (2019). Modelling acrual of arguments in ASPIC+. In *Proceedings of the 17th International Conference on Artificial Intelligence and the Law* (pp. 103–12).

Prakken, H. & Sartor, G. (1997). Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logic*, 1(2), 25–75.

Priest, G. (2002). Rational dilemmas. *Analysis*, 62(1), 11–6.

Pryor, J. (2000). The skeptic and the dogmatist. *Noûs*, 34, 517–49.

Pryor, J. (2018). The merits of incoherence. *Analytic Philosophy*, 59(1), 112–41.

Reiter, R. (1978). On closed world data bases. In H. Gallaire & J. Minker (Eds.), *Logic and Data Bases* (pp. 55–76). New York: Plenum.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligece*, 13, 81–132.

Rescher, N. & Manor, R. (1970). On inference from inconsistent premisses. *Theory and Decision*, 1:2, 179–217.

Ross, W. D. (1930). *The Right and the Good*. Oxford University Press.

Sartre, J.-P. (1996 [1946]). *L'existentialisme est un humanisme*. Folio Essais.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, T. M. (2000). Principles and particularisms. In *Proceedings of the Aristotelian Society, Suppl. Volume 74* (pp. 301–17).

Schechter, J. (2013). Rational self-doubt and the limits of closure. *Philosophical Studies*, 163(2), 429–52.

Schoenfield, M. (2014). Permission to believe: Why permissivism is true and what it tells us about irrelevant influences on belief. *Noûs*, 48(2), 193–218.

Schroeder, M. (2007). *Slaves of the Passions*. New York: Oxford University Press.

Schroeder, M. (2018). The unity of reasons. In D. Star (Ed.), *The Oxford Handbook of Reasons and Normativity* (pp. 46–66). Oxford University Press.

Searle, J. (1980). Prima facie obligations. In Z. van Straaten (Ed.), *Philosophical Subjects: Essays Prested to P. F. Strawson* (pp. 238–59). Oxford University Press.

Silva, P. (2017). How doxastic justification helps us solve the puzzle of misleading higher-order evidence. *Pacific Philosophical Quarterly*, 98(S1), 308–28. DOI: 10.1111/papq.12173.

Skipper, M. (2019). Higher-order defeat and the impossibility of self-misleading evidence. In M. Skipper & A. Steglich-Petersen (Eds.), *Higher-Order Evidence: New Essays* (pp. 189–208). Oxford University Press.

Sliwa, P. & Horowitz, S. (2015). Respecting all the evidence. *Philosophical Studies*, 172, 2835–58.

Tal, E. (f). Is higher-order evidence evidence? *Philosophical Studies*, forthcoming.

Thomason, R. (2018). Logic and artificial intelligence. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition.

Titelbaum, M. (2015). Rationality's fixed point (or: in defense of right reason). *Oxford Studies in Epistemology*, 5, 253–294.

van Fraassen, B. (1973). Values and the heart's command. *The Journal of Philosophy*, 70, 5–19.

Väyrenen, P. (2009). A theory of hedged moral principles. *Oxford Studies in Metaethics*, 4, 91–132.

Weatherson, B. (2013). Disagreements, philosophical, and otherwise. In D. Christensen & J. Lackey (Eds.), *The Epistemology of Disagreement: New Essays* (pp. 54–73). Oxford University Press.

Weatherson, B. (ms). Do judgments screen evidence? Unpublished manuscript, University of Michigan.

Wedgwood, R. (2010). The moral evil demons. In R. Feldman & T. Warfield (Eds.), *Disagreement* (pp. 216–46). Oxford University Press.

White, R. (2005). Epistemic permissiveness. *Philosophical Perspectives*, 19(1), 445–59.

White, R. (2007). Epistemic subjectivism. *Episteme*, 4, 115–29.

Worsnip, A. (2018). The conflict of evidence and coherence. *Philosophy and Phenomenological Research*, 96(1), 3–44.

Worsnip, A. (2019). Can your total evidence mislead about itself? In M. S. Rasmussen & A. Steglich-Petersen (Eds.), *Higher-Order Evidence: New Essays* (pp. 298–316). Oxford University Press.