This is the submitted version of the article forthcoming in *Pacific Philosophical Quarterly* (DOI: 10.1111/papq.12453). Please, cite the published version.

Epistemic conflicts and the form of epistemic rules

Aleks Knoks
University of Luxembourg

Abstract

While such epistemic rules as 'If you perceive that X, you ought to believe that X' and 'If you have outstanding testimony that X, you ought to believe that X' seem to be getting at important truths, it is easy to think of cases in which they come into conflict. To avoid classifying such cases as dilemmas, one can hold either that epistemic rules have built-in unless-clauses listing the circumstances under which they don't apply, or, alternatively, that epistemic rules are contributory. This paper explores both responses from a formal perspective, drawing on a simple defeasible logic framework.

Keywords: epistemic conflicts, epistemic rules, defeasibility, defeasible logic, particularism

Contents

1 Introduction

On a natural picture of rational belief, we aim to believe only what's true, we do our best to believe what's true by using the evidence we have, and we do so by relying on a set of *epistemic rules* that tell us—in a general way—what's most rational to believe under different epistemic circumstances. This picture is familiar, intuitive, and endorsed by many.[1] And even those who do not endorse it, or do not endorse it in full, may still think that epistemic rules, or something like them, have important roles to play.[2]

Now consider two sample epistemic rules:

(Perception) If an agent's epistemic situation includes a perception that X, the agent ought to believe that X.

(Testimony) If an agent's epistemic situation includes testimony that X, the agent ought to believe that X.[3]

[1] See (Boghossian 2008), Chisholm (1980), (Lasonen-Aarnio 2014), (Peacocke 2004), (Pollock & Cruz 1999, Ch. 5), and (Wedgwood 2002), among many others.

[2] Thus, those who hold that the goal of belief is something other than truth—e.g., knowledge, see Engel (2013)—might still think that our best shot at achieving it is to rely on epistemic rules. Evidentialists—who side with Conee & Feldman (2004) in holding that one ought to proportion one's beliefs to evidence—might try to spell out the way one actually proportions one's beliefs to evidence by appealing to epistemic rules. Foundationalists could appeal to epistemic rules to explain how justification is passed on from the secure foundation to non-foundational beliefs. Reliabilists can appeal to epistemic rules to spell out their notion of justification—see (Goldman 1986, Chs. 4–5). And even knowledge-firsters can put notions resembling epistemic rules to good use—see Lasonen-Aarnio's (2010) discussion of "good methods".

[3] Both rules have enjoyed much attention in recent epistemology literature—see, e.g., (Chisholm 1980), (Huemer 2000), (Pollock 1995, 2001), and (Pryor 2000) for Perception and (Bradley 2019), (Elga 2007, 2010), and (Titelbaum 2015) for Testimony. One might wonder about the meaning of the term *epistemic situation* in their statement. It's natural to think that the correct normative theory will 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,

Notice that anyone who thinks that these rules are genuine must be able to explain what happens in cases where they support jointly inconsistent recommendations and in cases where they get undermined: suppose that you see a red-looking object, while an extremely reliable source tells you that this object is blue. Or suppose that your greatest authority in epistemology tells you that Testimony is false. What are you to do in these cases? Applying the above rules in the first quickly leads to the conclusion that you ought to believe that the object is red, and that it is blue! And applying Testimony in the second leads to the conclusion that you ought to disbelieve it—which is at the very least odd, if we think that Testimony is a genuine rule.[4] Call this the *problem of epistemic conflicts*.[5]

There are two plausible strategies of response to this problem, and both weaken the above statement of the rules, or suggest that they are *defeasible*.[6] According to the first, Perception, Testimony, and other rules have implicit hedges or unless-clauses listing the circumstances under which the rule doesn't apply. Adopting this strategy, one can contend 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 the above rules has to be changed for "has a reason"—the thought here is that rules (by themselves) do not specify what doxastic attitudes you ought to have, but only

facts about her condition, her past, or other kinds of facts. The totality of all of these normatively-relevant descriptive features is the agent's epistemic situation—cf. (Titelbaum 2015, Sec. 2).

[4]he second example is actually a special case of a situation where a rule gets undermined, namely, one where a rule self-undermines. Also, the conclusion I draw relies on a(n inter-level coherence) principle, saying that it's never epistemically permissible to believe that you ought to disbelieve X and believe X all the same. Given the goals in this paper, nothing of substance hinges on presupposing this principle.

[5]My formulation of the problem follows Bradley's (2019).

[6]I'm using the term *defeasible* loosely at this point, meaning that a rule can engender an ought in one situation and then fail to engender an ought in a situation that is only a little different from the original one. We'll get more precise on this term as we proceed.

what counts in favor or against having them. Adopting this second strategy, one can contend that there's indeed a conflict between Perception and Testimony, but that it is resolvable, because, say, your perception outweighs the testimony.[7]

We will state the views on rules that come with each response precisely later. For now let's just label them, respectively, the *hedged-rules view* and the *contributory-rules view*, and note that they are naturally thought of and typically presented as distinct.[8]

This paper has two goals. The first is to express the problem of epistemic conflicts in a formal framework and model the views on rules that come with each response—here I draw on ideas from the defeasible logic paradigm and, more specifically, default logic.[9] There will be three models: one captures the contributory-rules view, the other two different versions of the hedged-rules view. My second goal is to establish a type of correspondence result between two of these models and explore its implications. My hope is that reaching these goals will contribute to our understanding of epistemic rules and their defeasibility.

The bulk of the paper is structured as follows. Section 2 presents the basic concepts and the problem formally. Sections 3 develops a model of the contributory-rules view. Section 4 and 5 develop models of two different versions of the hedged-rules view: on the first, the content of rule

[7]The authors who have pursued the first strategy include Bradley (2019), Elga (2010), and Titelbaum (2015), in epistemology, and Holton (2002) and T. M. Scanlon (2000), in ethics. The second strategy is really at home in ethics where it is associated with W. D. Ross (1930). In epistemology, Bradley (2019) ascribes it to Christensen (2007, 2010, 2013). There are also authors whose views on rules seem to combine both strategies, such as Horty (2012) and Pollock (1995, 1999). As far as I see, the two strategies (and their combination) exhaust the space of plausible responses to the problem that don't do away with rules as such. In this paper, I assume that both strategies succeed in showing that epistemic conflicts are not *tragic dilemmas*, or situations in which the agent ends up failing to be as she ought (epistemically) no matter what she does. Given my goals, the assumption seems innocuous. I will later (re)state it in a mathematically precise way—see footnotes 15, 20, and 28.

[8]See, e.g., (Bradley 2019) in epistemology and (Dancy 2004, Sec. 1.2) in ethics. I adopt the labels from Bradley.

[9]See (Reiter 1980), as well as (Horty 2012) and (Makinson 2005, Ch. 4). I use Horty's user-friendly notation.

hedges is descriptive; on the second, it is normative, referring to beliefs the agent can justifiably hold. Section 6 establishes the central result: the model of the contributory-rules view turns out to be equivalent to a *fragment* of the first model of the hedged-rules view. Section 7 exhibits one implication of this result in a discussion of the expressive power of the three models and their prospects of capturing the familiar distinction between rebutting and undercutting defeat. This is followed by a discussion of some broader implications of our formal exploration in Section 8 and an appendix where the main observations are verified.

A note before we proceed: the models presented here may not capture every possible way of thinking about contributory and hedged rules. So the results I obtain and the conclusions I draw come with the qualification "for the versions of views that are captured by the models". I do think, however, that the models capture the cores of the views adequately, putting the views themselves in sharper focus, and I take great care to set the models up in a stepwise fashion to convince the reader that they do.[10]

[10]As any models, the ones set up here rest on some simplifying assumptions. Perhaps most importantly, in them, the epistemic rules are what we might call *flattened*. Authors writing on epistemic rules often distinguish between rules that license formation of beliefs based on the occurrence of nondoxastic states or events such as perceptions, on the one hand, and rules that license formation of beliefs based on logical, inductive, or abductive relations to other beliefs. Borrowing the terminology from Goldman (2009), we can call them, respectively, *noninferential* and *inferential rules*. The models discussed in this paper do not allow for explicit representation of inferential rules. Instead, they allow for explicit representation of noninferential rules, as well as proxies representing chains of rules that start with noninferential ones—this will be discussed in more details in Section 5. I think of this as a useful simplification that lets us focus on the question of what the form of epistemic rules has to be if the problem of epistemic conflicts is to have a solution.

2 Preliminaries

As our background, we assume the language of propositional logic with the standard connectives. The turnstile $\vdash$ will stand for classical logical consequence: 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 to avoid unnecessary clutter when formalizing particular cases, we assume that our background language allows for materially inconsistent atomic formulas that cannot jointly be true, formulas expressing such propositions as "The object in front of you is red" and "The object in front of you is blue." Also, to have a more natural way of stating Perception and Testimony, we extend the language with three designated predicates $Perceive(\cdot)$, $Testimony(\cdot)$, and $Believe(\cdot)$. The formula $Perceive(X)$ captures the idea that the agent perceives that X; $Testimony(X)$ that the agent has testimony that X; and $Believe(X)$ that the agent believes that X. The reference to an agent is important: all formulas should be thought of as relativized to an agent in some epistemic situation. We also make use of the customary deontic operator $\bigcirc(\cdot)$. A formula of the form $\bigcirc Believe(X)$, then, says that the agent ought to believe that X. Note that the sense of *ought* that I have in mind here is epistemic, as opposed to pragmatic or any other sense, and that it is all things considered, as opposed to pro tanto.

One might be tempted to express Perception and Testimony—as well as other rules—as material conditionals, that is, as, respectively, $Perceive(X) \supset \bigcirc Believe(X)$ and $Testimony(X) \supset \bigcirc Believe(X)$. However, given that we'll eventually want to make these rules defeasible, it will be more convenient to start off thinking about them by analogy with the (indefeasible) inference rules of logical systems.

The rule of conjunction elimination sanctions concluding X (and Y) whenever one has been able to establish $X\&Y$:

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

Similarly, we can think of epistemic rules as sanctioning drawing a certain type of conclusion whenever one's epistemic situation includes a certain type of feature:

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

It's worth highlighting an important detail: what we see here—in the case of conjunction elimination, as well as Perception and Testimony—are actually not rules, but *rule schemas*. It's common to elliptically refer to rule schemas as *rules*, but the two notions are distinct, and shouldn't be confused. We denote rules using the letter r, with subscripts.

Let us also introduce two functions $Premise[\cdot]$ and $Conclusion[\cdot]$, for picking out the premise and conclusion of a given rule. Thus, where r stands for the rule $\frac{X}{Y}$, the expression $Premise[r]$ stands for the proposition X and $Conclusion[r]$ for the proposition Y. We'll apply the second function to sets of rules too: 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 represent epistemic situations as pairs of the form $\langle \mathcal{W}, \mathcal{R} \rangle$, refer to them as *contexts*, and denote them using the lowercase c, with subscripts. The first element of a context—its *hard information*—is just a set of ordinary propositional formulas, expressing the descriptive features of the epistemic situation. And the second element $\mathcal{R}$ is a set of epistemic rules, that is, rules of the form $\frac{X}{\bigcirc Believe(Y)}$. As an illustration, consider the case where you're looking at an object in front of you and it looks red to you. Letting R stand for the proposition that the object is red, we can express this situation as the context $c_1 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ contains the formula $Perceive(R)$, saying that the object looks red to you, and where $\mathcal{R}$ contains the rule $r_1 = \frac{Perceive(R)}{\bigcirc Believe(R)}$, saying that you ought to believe that the object is red in case it looks red to you.

Why do we have the intuition that Perception and Testimony reveal something important? A good first-pass answer is that their instances—together with other rules—are what link the descriptive features of epistemic situations to the normative ones.[11] The formal notion of a contexts

[11] Compare to Bradley's (2019) suggestion that one can take epistemic rules to be grounding principles.

suggests that we explicate this intuitive idea as follows. There's a context standing for every epistemic situation, and the (infinite) set of all contexts shares a common set of rules $\mathcal{R}$, containing every instance of Perception, Testimony, and other rule schemas.

One might hope that the logic governing the operation of epistemic rules is just the plain classical logic. This idea can be captured in our framework in two simple steps. The first introduces the notion of *triggered rules*:

- Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a 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 from $\mathcal{R}$ whose premises can be derived from $\mathcal{W}$ by classical logic. It's easy to see that r_1 is triggered in c_1: since $Perceive(R)$ is in $\mathcal{W}$ and $Premise[r_1] = Perceive(R)$, we have $\mathcal{W} \vdash Perceive(R)$. In the second step, we specify which ought-formulas follow from a context, on the basis of triggered rules. There are a few ways one could proceed here. We adopt the most straightforward one, simply collecting the conclusions of all rules that are triggered:[12]

- Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context. Then the ought statement $\bigcirc X$ follows from c just in case $\bigcirc X \in Conclusion[Triggered_c]$.

It's easy to verify that $\bigcirc Believe(R)$ follows from c_1, on this definition. Since r_1 is the only rule triggered in c_1, the set $Conclusion[Triggered_{c_1}]$ equals $\{\bigcirc Believe(R)\}$.

Now recall the troubling situation we started with, the one where an object in front you looks red and a reliable source—say, your friend Walter—tells you that it's blue. Let R be as before

[12] I've chosen this definition to keep the formalism as simple as possible. One of its drawbacks is that, on it, the formulas $\bigcirc A$ and $\bigcirc B$ might follow from some context c without $\bigcirc(A \& B)$ following from c too. But nothing important hinges on this: we can substitute the definition with a more involved one—the formula $\bigcirc X$ follows from $c = \langle \mathcal{W}, \mathcal{R} \rangle$ if and only if $\bigcirc X$ follows from $Conclusion[Triggered_c]$ by *standard deontic logic*—and retain all results.

and let B stand for the proposition that the object in front of you is blue. This situation can be encoded in the context $c_2 = \langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W}$ containing $Perceive(R)$ and $Testimony_w(B)$, and $\mathcal{R}$ containing the familiar rule $r_1 = \dfrac{Perceive(R)}{\bigcirc Believe(R)}$ and the new rule $r_2 = \dfrac{Testimony_w(B)}{\bigcirc Believe(B)}$, saying that you ought to believe that the object is blue if you have testimony that it is. 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, and it's clear that we need to change either the way we think about rules, the logic responsible for generating ought-formulas, or both.

3 Contributory rules

One response to the problem suggests that Perception and Testimony don't really say that you ought to believe that X in any situation where you have, respectively, perception that X and testimony that X, but only that you *have a reason* to believe that X. How does this help? Well, suppose you're in the above case. If the two rules are contributory, we only get the conclusion that you have a reason to believe that the object is red and a reason to believe that it's blue. And given that reasons are pro tanto, a dilemma gets reduced to a conflict we can live with. This section explores what the view on rules that comes with this response looks like in our formal framework.

One might be tempted to express Contributory Testimony and Perception as, respectively,

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

And if all one cares about is to block the problem, these schemas do the job. For "There's a reason to $Believe(X)$" will never be in contradiction with "There's a reason to $Believe(Y)$," even if X and Y are inconsistent. However, there's good reason to be dissatisfied with this. Recall that rules were supposed to get us from a description of any particular epistemic situation to the doxastic attitudes the agent ought to have. In our formal setting, this is the question of which ought-statements follow

from a given context. The problem is that it's unclear how we can get to any ought-statement from "There's a reason to $Believe(X)$" and "There's a reason to $Believe(Y)$."

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

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

How are their instances to be interpreted? Let's take $r_3 = \dfrac{Testimony_w(R)}{Believe(R)}$ as an example. Intuitively, it says that $Testimony_w(R)$ *counts in favor of* believing R, or that Walter's testimony that the object is red counts in favor of believing that it's red. Functionally, r_3 lets us infer $Believe(R)$ from $Testimony_w(R)$ *by default*. The qualification is important. It's added because the presence of $Testimony_w(R)$ doesn't guarantee that it will be possible to infer $Believe(R)$, as other rules might come in the way. (How this can happen will become clear in a minute.) A major advantage of expressing contributory rules in this way is that there's a method for deriving ought-formulas from them within a hand's reach.

It's both natural and standard to associate contributory rules with relative weights. To represent them in the model, we introduce a new device, a *priority relation* over rules.[13] Where r and r' are (contributory) rules, a statement of the form $r \leq r'$ means that r' has at least as much weight as r does, or that r' is at least as strong as r. We require that the relation $\leq$ satisfies some natural properties. First, it must satisfy *reflexivity*,

$$r \leq r,$$

saying that each rule is at least as strong as itself. And second, it must satisfy *transitivity*,

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

[13]The move is standard—see, e.g., (Horty 2012) and (Pollock 1995).

saying that 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. It'll be useful to introduce some shorthand: when we have $r \leq r'$ without $r' \leq r$, we write $r < r'$.

Above, we expressed epistemic situations as context. Now we express them using three-element structures I call *weighted contexts*. In addition to the hard information $\mathcal{W}$ and a set of contributory rules $\mathcal{R}$, these include a priority relation on the rules in $\mathcal{R}$. Our first example of a weighted context $c_3 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ expresses the problematic case where an object looks red to you, but Walter tells you that it's blue. Its hard information $\mathcal{W}$ is comprised of $Preceive(R)$ and $Testimony_w(B)$, and its set of rules $\mathcal{R}$ includes $r_3 = \dfrac{Testimony_w(B)}{Believe(B)}$ and $r_4 = \dfrac{Perceive(R)}{Believe(B)}$, with the latter having strictly more weight than the former, $r_3 < r_4$.

Having defined weighted contexts, we will specify which ought-formulas follow from them. As a first step, we introduce the notion of *contrary rules*:

- Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a weighted context and $r = \dfrac{X}{Believe(Y)}$ and $r' = \dfrac{Z}{Believe(Y')}$ two rules from $\mathcal{R}$. Then r and r' are *contrary in the context* c, written as $contrary(r, r')$, just in case Y and Y' are inconsistent.[14]

Notice that this definition qualifies r_3 and r_4 as contrary in the context c_3.

As our next step, we define the notion of *binding rules*. Intuitively, these are the rules that are triggered in the context and that have more weight than all the rules contrary to them. The formal definition runs thus:

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

[14]This definition works under the simplifying assumption that conflicts between rules are always conflicts between two rules. Nothing important hinges on this, however: while there can be sets of rules the conclusions of which are pairwise consistent, but jointly inconsistent, it's also possible to generalize the definition, taking care of such sets. Also, in Section 7, we will talk about contrary *hedged* rules. The definition extends to them straightforwardly.

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

So, for a rule to qualify as binding, it has to be triggered and it can't be *counterbalanced*, that is, there can't be another rule that's also triggered, contrary to it, and that has at least as much weight. Applying this definition to c_3, we can verify that r_3 isn't binding in c: while it gets triggered, there's the rule r_4 that also gets triggered, has more weight than r_3, and is contrary to r_3. The rule r_4, by contrast, is binding in c_3.

With this, we can specify when ought-formulas follow from weighted contexts:

- Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be a weighted context. Then the statement $\bigcirc X$ follows from c just in case $X \in Conclusion[Binding_c]$.

Notice the two changes from the first-pass definition from Section 2. First, $Binding_c$ has been substituted for $Triggered_c$. Second, X has been changed for $\bigcirc X$ in the final expression, reflecting the fact that we're dealing with contributory rules. Pushing on with our example, it's easy to verify that the set $Conclusion[Binding_{c_3}]$ equals $\{Conclusion[r_4]\} = \{Believe(R)\}$, implying that $\bigcirc Believe(R)$ follows from c_3, while $\bigcirc Believe(B)$ does not. There's nothing dilemmic.

So far, so good. But part of the problem of epistemic conflicts is that we can always think of variations in a given epistemic situation that prompt changes in the normative landscape. Thus, suppose that your situation includes a further testimony that the object is blue, and that it comes from your extraordinarily reliable friend Sue. Suppose further that in relevantly similar cases in the past Sue has turned out to be right about a given object's color considerably more often than your perception. Now it seems that it's more reasonable for you to believe that the object is blue—or so I will assume. How can our formalized version of the contributory-rules view handle this? Well, I propose to capture the new scenario in the context $c_4 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, where the hard information $\mathcal{W}$

is comprised of $Perceive(R)$, $Testimony_w(B)$, and the new formula $Testimony_s(B)$—which says that your epistemic situation includes Sue's outstanding testimony—where the set of rules $\mathcal{R}$ includes the familiar r_3, r_4, along with the new $r_5 = \dfrac{Perceive(B)}{Believe(B)}$; and where the ordering on rules is $r_3 < r_4 < r_5$. It's not difficult to verify that the context c_4 entails $\bigcirc Believe(B)$, and that it doesn't entail $\bigcirc Believe(R)$. Again, there's nothing dilemmic here, and it should be clear that further variations of the case can be handled in a similar fashion. In fact, any conflict between contrary epistemic rules is guaranteed to get resolved, if we assume—which we do—that they are connected by the priority relation.[15]

I will have more to say about the expressive power of the model we just set up in Section 7, but, for now, we have all we need and can turn to the hedged-rules view.

4 Hedges listing descriptive features

According to the second response to the problem of epistemic conflicts, rules have built-in hedges. This general suggestion can be taken in two different directions. According to the first, the content of rule hedges is *descriptive*; according to the second, it is *normative*. In this section, we explore the first direction; in the next, the second one.

To see the basic idea at work, recall the case where an object looks red to you, but your friend Walter tells you that it is blue. Applying Perception and Testimony leads to the conclusion that you ought to believe that the object is red and that it is blue. But now suppose that we supplement Testimony with a hedge, as follows:

[15] In footnote 7, I already stated my assumption that the move to contributory rules succeeds in showing that cases involving conflicts between rules are not tragic dilemmas. In our formal framework this assumption amounts to a restriction on the class of weighted contexts: for any weighted context $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$, we require that, for any two rules r, r' in $\mathcal{R}$ such that $contrary(r, r')$, either $r \leq r'$, or $r' \leq r$.

(Hedged Testimony) If an agent's epistemic situation includes testimony that X, the agent ought to believe that X—unless the situation also includes a perception that's contrary to X.

Notice that Hedged Testimony doesn't issue a recommendation in the scenario. Even though you have testimony that the object is blue, it's not the case that you ought to believe that it's blue, since you perceive it as looking red. By contrast, Perception—assuming we leave it as is—does entail that you ought to believe that the object is red.

Now let's express this in our framework. First off, we need to modify the original notion of a rule. What we're going to do is transform the schema $\frac{X}{\bigcirc Believe(Y)}$, familiar from Section 2, into the slightly more complex $\frac{X : \neg Z_1, \ldots, \neg Z_n}{Believe(Y)}$. The new element $\{\neg Z_1, \ldots, \neg Z_n\}$, which we'll occasionally abbreviate as $\mathcal{Z}$, corresponds to the rule's hedge, and, formally, it's just a set of negated propositional formulas. A hedged rule of this form should be read as, "If X obtains, then conclude $Believe(Y)$, unless either Z_1, or Z_2, …, or Z_n obtain", or, alternatively, as, "If X obtains, and it can be assumed that not-Z_1, not-Z_2, …, and not-Z_n, then $Believe(Y)$."[16]

We keep the functions for selecting rule premises and conclusions. In addition, we introduce a new function selecting rule hedges (if any): if the rule r is of the form $\frac{X : \neg Z_1, \ldots, \neg Z_n}{Y}$, then $Hedge[r]$ is the set $\{\neg Z_1, \ldots, \neg Z_n\}$, and if r is of the form $\frac{X}{Y}$, then $Hedge[r]$ is the empty set.

Above, we expressed epistemic situations using (weighted) contexts. Now we'll express them as *hedged contexts*. These are pairs of the form $\langle \mathcal{W}, \mathcal{R} \rangle$, just like plain contexts, but their second element $\mathcal{R}$ can contain hedged rules.[17] Our first example of a hedged contexts is meant to capture

[16] Compare to Reiter's (1980) default rules.

[17] Some technical notes: it's in principle possible for a context to contain two rules whose premises and conclusions are the same, but hedges different. Intuitively, however, such pairs are deviant: instead of two rules, there should be only one. So we assume that hedged contexts never contain such deviant pairs of rules, or that, for any $r, r' \in \mathcal{R}$, in case $Premise[r] = Premise[r']$ and $Conclusion[r] = Conclusion[r']$, then $r = r'$. Similarly, we assume that

the case at hand. Consider $c_5 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ consists of $Perceive(R)$ and $Testimony_w(B)$, while $\mathcal{R}$ contains the rules

$$r_4 = \frac{Perceive(R)}{Believe(R)} \text{ and}$$

$$r_6 = \frac{Testimony_w(B) : \neg Perceive(R)}{Believe(B)}.$$

Now we need to specify a procedure for obtaining ought-formulas from hedged contexts—in particular, it should let us derive $\bigcirc Believe(R)$ from c_5. As a first step, we introduce the notion of *admissible rules*—which will play a similar role to that of binding rules from the previous section:

- Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a hedged context and $\mathcal{R}'$ a set of rules from $\mathcal{R}$, that is, $\mathcal{R}' \subseteq \mathcal{R}$. Then the rules from $\mathcal{R}$ that are *admissible* relative to $\mathcal{R}'$, against the background of c, are those that belong to the set

$$Admissible_c(\mathcal{R}') = \{r \in \mathcal{R} : r \in Triggered_c \text{ and, for no } \neg Z \in Hedge[r], \mathcal{W} \cup Conclusion[\mathcal{R}'] \vdash Z\}.$$

Notice that admissible rules (unlike triggered and binding ones) are defined relative to a subset of rules from the context. The definition itself, then, says that, of all the rules in $\mathcal{R}$, only those are admissible—relative to $\mathcal{R}'$—that are triggered and also such that none of the formulas listed in their hedges follow from the hard information and the conclusions of rules from $\mathcal{R}'$.[18]

To see this definition at work, we apply it to c_5. Relative to $\{r_4, r_6\}$, the set of admissible rules, $Admissible_{c_5}(\{r_4, r_6\})$, is the singleton $\{r_4\}$: even though r_6 is triggered, the formula $Perceive(R)$ that is listed in its hedge follows from the context's hard information, $\mathcal{W} \vdash Perceive(R)$. And relative to the empty set, $\varnothing \subset \mathcal{R}$, the set of admissible rules, $Admissible_{c_5}(\varnothing)$, is again $\{r_4\}$, for parallel reasons.

The next notion we define is that of a *stable set* of rules:

contexts do not contain any deviant rules that list their own premises in their hedges, or that there's no $r \in \mathcal{R}$ such that $\neg Premise[r]$ in $Hedge[r]$.

[18]The need to relativize the notion of admissible rules will not become apparent before the next section. If we only cared about the model set up in this section, the reference to a subset of rules would be superfluous.

- Given a hedged context $c = \langle \mathcal{W}, \mathcal{R} \rangle$, we say that a set of rules $\mathcal{R}' \subseteq \mathcal{R}$ is *stable* in c just in case $\mathcal{R}' = Admissible_c(\mathcal{R}')$.

So a subset of rules is stable just in case it contains all and only those rules that are admissible relative to it. It's not difficult to see that neither $\{r_4, r_6\}$, nor $\varnothing$ qualify as stable—the former contains the inadmissible rule r_6, while the latter doesn't contain the admissible r_4—and that the only stable scenario based on context c_5 is $\{r_4\}$.

Having defined stable sets of rules, we specify the conditions under which an ought-statement follows from a hedged context:

- Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a hedged context. Then $\bigcirc X$ follows from c if and only if X is in $Conclusion[R']$ for *some* stable set of rules $\mathcal{R}'$ based on c.[19]

Since $\{r_4\}$ is the only stable set based on c_5 and $Conclusion[\{r_4\}] = \{Believe(R)\}$, we get the desired result that $\bigcirc Believe(R)$ follows from c_5.

With this, we have all the formal definitions we need, and can turn to the more philosophically interesting question: how can our formal version of the hedged-rules view handle further variations in the epistemic situation? Recall the case from Section 2 in which your extraordinarily reliable friend Sue tells you that the object is blue. The most straightforward way to capture it is as the context $c_6 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ is comprised of $Perceive(R)$, $Testimony_w(B)$, and $Testimony_s(B)$, and where $\mathcal{R}$ contains the rules

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

[19] A reader familiar with the work of Horty (2012)—or the defeasible logic literature this work draws on—will have noticed that this definition adopts the *conflict*, as opposed to the *disjunctive account*. Such a reader may wonder how I might justify this choice. The short answer is that I don't think that anything important hinges on which account we choose here—since we will focus on contexts that have only one stable set—and that the conflict account is more permitting with respect to the existence of dilemmas, and, thus, also more fitting in the present context.

$r_6 = \dfrac{Testimony_w(B) : \neg Perceive(R)}{Believe(B)}$, and

$r_7 = \dfrac{Testimony_s(B)}{Believe(B)}$.

It's not difficult to verify that $\bigcirc Believe(B)$ follows from c_6, which seems intuitively correct. And it shouldn't be difficult to see that further variations of the case can be handled in a similar fashion. In fact, any conflict between two contrary epistemic rules r and r' can be resolved by adding a formula of the form $\neg Premise[r]$ to the hedge of r'.

This way of dealing with situations involving conflicts between rules avoids classifying them as dilemmas.[20] However, it also gives rise to a worry: it seems that the complexity of rule hedges can spiral out of control, and that the view itself may turn out to reduce to a version of particularism in epistemology, or, roughly, the view that there either are no epistemic principles, or that such principles play only a marginal role.[21] Here it pays to recall the distinction between rules and rule schemas. In our framework, the question of whether a view on rules is particularist boils down to the question of whether one can reasonably hold that rule schemas are prior to rules. Let's use the Perception schema as illustration. Starting with its instances—hedged rules we would use in contexts representing various epistemic situations we might imagine—we can ask whether the number of *types* of (descriptive) features that occur in their hedges is finite, and whether or not these features are always present. If so, we'll be able to write down an informative schema covering all of them, that is, a schema of the form $\dfrac{Perceive(X)\text{: } \neg Feature1(Y), \ldots, \neg FeatureN(Y)}{Believe(X)}$, where the background assumption is that it's not rational for the agent to believe both X and Y. If such a schema can be written down, we could hold that it is prior to rule instances, that it is

[20] Recall that we are assuming that the move to hedged rules is enough to avoid classifying epistemic conflicts as dilemmas. In the present context, the assumption amounts to a restriction on the class of acceptable hedged contexts: for any hedged context $c = \langle \mathcal{W}, \mathcal{R} \rangle$, we require that, for any two rules r, r' in $\mathcal{R}$ such that $contrary(r, r')$, either $\neg Premise[r] \subseteq Hedge[r]$, or $\neg Premise[r'] \subseteq Hedge[r]$.

[21] Cf. (Bradley 2019, p. 12ff).

what gives rise to them, and that it is a genuine epistemic principle, of the sort particularists claim don't exist. If, however, it turned out that there are infinitely many *types* of features that occur in the hedges of rules, or they don't always occur in them, then we won't be able to write down a corresponding schema that would both do justice to all rules and be informative. For were we to opt for informativeness, taking into account every rules, we couldn't write the schema down. And were we to opt for something that can be written down, we'd end up with something like "If an agent's epistemic situations includes a perception that X, then she ought to believe that X—unless something comes in the way". In such a case, I think, we'd have to say that rule instances are prior to this uninformative schema, and to concede that Perception is a generalization and not a genuine epistemic principle.

I submit that the prospects for an informative Hedged Perception schema are rather dim. First off, it seems plausible that we will need to list numerous features in its hedge to account for the various cases discussed in the literature—just think of red lights, drugs that make blue things look red, situations where your perception has systematically mislead you in the past. What's more, we will also need to account for cases in which further features preclude the sabotage of Perception—thus, you might be told that the information about red lights is a lie, that you have been given an antidote, or that the current situation is radically different from the seemingly similar past ones—as well as cases in which the normative effect of these further features itself gets sabotaged, an so on. In fact, the two simple cases we discussed already exhibit the looming complexity: since we want the instance of Hedged Perception to apply in the first, but not the second one, we cannot identify $Feature1(Y)$ with "the agent has testimony that Y". Instead, it would have to be something like "the agent has testimony that Y, with the testifier having such and such track record". This already looks fairly complicated, casting doubt on the usefulness of the rule.

The overall conclusion that invites itself is that the view on which rule hedges list descriptive

features avoids the problem of epistemic conflicts, but also likely reduces to a particularist view.

5 Hedges with normative contents

The second way of thinking about hedged rules holds that the contents of their hedges is normative. This section sketches two version of this view and develops a formal model of the one that holds more promise in resolving the problem of epistemic conflicts.

In epistemology, the authors who have explored normative versions of the hedged-rules view include Elga (2010), Titelbaum (2015), and, possibly, Bradley (2019)—Bradley's view is ambiguous and may well be closer to the descriptive view discussed above. Elga and Titlebaum both hold that hedges of *genuine* epistemic rules make them immune to counterevidence. Thus, Titelbaum's preferred version of Testimony runs, roughly, as follows: if an agent's situation includes testimony that X, then the agent ought to believe that X—unless X contradicts an *a priori* truth about what rationality requires.[22] Notice that, assuming that rationality always requires one to believe this version of Testimony, any evidence suggesting that it is false can be discounted, making the rule immune to undermining. And Titelbaum does provide some ingenious arguments to the conclusion that one always has *a priori* justification for believing his preferred version of Testimony.

While this view is very interesting, I don't think it will do for our purposes. It provides a way out of cases where an agent receives evidence that Testimony is false, but it seems to be of little help in resolving cases in which epistemic rules require having inconsistent beliefs. So, I won't pursue it any further.

A different take on the contents of hedges can be gleaned from Bradley. His version of Testimony runs thus:

[22] See (Titelbaum 2015, p. 274).

(Hedged Testimony*) If an agent's epistemic situation includes testimony that X, then the agent ought to believe that X—unless the agent has evidence that is inadmissible-relative-to-Testimony.[23]

On my reading, the "inadmissible evidence" that this schema refers to makes it justifiable for the agent to believe that the schema is not safe to rely on in the given circumstances.[24] Notice that this idea provides for a plausible explanation of what is going on in the case where you receive the unexpected testimony from Sue. Given Sue's track record of being more reliable than your own perception, it seems reasonable for you to conclude that you must be in the sorts of circumstances where Perception is not safe to rely on and to go with Testimony instead.

I think that this way of fleshing out the view is natural, and that it holds promise in resolving the problem of epistemic conflicts. So let's express it in our formal framework.

The first thing we need to do is extend our background language in two ways. First, we introduce rule names, assigning every (hedged) rule a unique name—and we use the letter r here, with subscripts. Second, we introduce a designated predict $Out(\cdot)$ that takes rule names as argument. The intended meaning of a formula of the form $Out(r)$ then is that r is not safe to rely on.[25]

With this, we can express the Hedged Testimony* and Perception* schemas as follows:

$$r(X) = \frac{Testimony(X) : \neg Believe(Out(r))}{Believe(X)},$$

$$r'(X) = \frac{Perceive(X) : \neg Believe(Out(r'))}{Believe(X)}.$$

[23] See (Bradley 2019, p. 11).

[24] Bradley seems to want Hedged Testimony* to refer to evidence that's inadmissible with respect to (unhedged) Testimony, and then have a further rule that refers to evidence that's inadmissible with respect to Hedged Testimony*—see (Bradley 2019, Sec. 7). I have to admit that I couldn't find a way to express this suggestion formally, and so I don't fully understand how it is meant to work.

[25] I borrow the idea from Horty (2012). Unlike him, however, I use the same name to refer to rules in our extended propositional language as I do when referring to them in English.

Note that these two schemas exhibit the general shape of any rule in our model of the view: any given rule r will contain a formula of the form $\neg Believe(Out(r))$ in its hedge, having the power to make this rule inadmissible. This seems well in line with Bradley's claims that all epistemic rules are hedged, and that there are no indefeasible rules. Also, in our model, the hedge of any rule r will contain *only* the formula $\neg Believe(Out(r))$.[26] One clear benefit of this is that the worries associated with particularism—discussed in the previous section—do not apply, at least, not immediately.

Now recall our two running examples: in the first, there's a conflict between your perception and Walter's testimony; in the second, between your perception and Walter's and Sue's testimonies. The first case can be captured in the context $c_7 = \langle \mathcal{W}, \mathcal{R} \rangle$, where $\mathcal{W}$ is comprised of $Perceive(R)$ and $Testimony_w(B)$, while $\mathcal{R}$ consists of the rules

$$r_7 = \frac{Perceive(R)\colon \neg Believe(Out(r_7))}{Believe(R)},$$

$$r_8 = \frac{Testimony_w(B) : \neg Believe(Out(r_8))}{Believe(B)}, \text{ and}$$

$$r_9 = \frac{Perceive(R) \& Testimony_w(B) : \neg Believe(Out(r_9))}{Believe(Out(r_8))}.$$

The second case, in turn, can be captured in the context $c_8 = \langle \mathcal{W}', \mathcal{R}' \rangle$, where $\mathcal{W}'$ extends $\mathcal{W}$ with the formula $Testimony_s(B)$ and $\mathcal{R}'$ extends $\mathcal{R}$ with the rules

$$r_{10} = \frac{Testimony_s(B) : \neg Believe(Out(r_{10}))}{Believe(B)},$$

$$r_{11} = \frac{Perceive(R) \& Testimony_s(B)\colon \neg Believe(Out(r_{11}))}{Believe(Out(r_7))}, \text{ and}$$

[26] This is a departure from Bradley's discussion. He seems to think that referring to "evidence that is inadmissible-relative-to-a-rule" in a hedge commits one to a view on which lists of exceptions to rules are long and possibly even not finitely statable—see (Bradley 2019, p. 12ff). That's why I think that Bradley's take on hedged rules is ambiguous between the sort of view sketched in this section and the one explored in the previous one. I take it to be an advantage of using formal tools in the present context that they help us clearly see that the two views are different.

$$r_{12} = \frac{Perceive(R)\&Testimony_w(B)\&Testimony_s(B) : \neg Believe(Out(r_{12}))}{Believe(Out(r_9))}.$$

Notice that r_7, r_8, and r_{10} are instances of Hedged Testimony* and Perception*. But what about r_9, r_{11}, and r_{12}? How do we make sense of them? Let's take r_9 as an example—parallel considerations apply to r_{11} and r_{12}. Those sympathetic to the picture on which justification for our beliefs derives from epistemic rules would say that the justification for the agent's conclusion that Hedged Testimony* is not safe to rely on—$Out(r_8)$ in our notation—must derive from several epistemic rules that we haven't discussed in this paper. For instance, they could appeal to a rule that justifies forming beliefs about seeming, a rule that puts the agent in the position to justifiably believe that she perceived the object as red-looking and that Walter says that this object is blue. Further, the could say that there's another rule corresponding to the inference to the best explanation that justifies the agent's forming the belief that Walter is mistaken on the basis of these two beliefs—or, perhaps, these two beliefs and the relation that obtains between the propositions involved.[27] The right way to think of r_9, then, is as flattening and compressing a chain of instances of the epistemic rules—whatever they happen to be—that justify the agent's belief that Testimony* is not safe to rely on in the situation at hand. Thus, in the present context, r_9, r_{11}, and r_{12} can be thought of as useful simplifications that do not smuggle anything that's foreign to the view that we are trying to express formally—or so I hope.

It's not difficult to verify that $\bigcirc Believe(R)$ follows from c_7, and that $\bigcirc Believe(B)$ follows from c_8, as desired. And it should also be apparent that the model we just set up has the resources to (re)describe any case involving a conflict between epistemic rules as a context where at least one of the rules doesn't apply and the agent can justifiably believe that it doesn't.[28]

[27] See (Goldman 2009, Sec. 2) for a recent suggestion that there may be an epistemic rule capturing inference to the best explanation—see also (Christensen 2010, p. 192ff).

[28] In the model, the intuitive idea that an agent can justifiably believe that the rule r doesn't apply corresponds to the derivability of $\bigcirc Believe(Out(r))$ from the context. So, the assumption that rule hedges with normative contents

6 A correspondence result

Now we have three models, regimenting different pictures of epistemic rules and the way conflicts between them get resolved. According to the first, rules are contributory, specifying what counts in favor or against the agent's beliefs. When such rules come in conflict, the rule that has more weight wins out and the one that has less gets defeated. The remaining two pictures share the ideas that rules have built-in hedges specifying the conditions under which they don't apply, and that, whenever two rules come into conflict, one of them doesn't apply because some conditions specified in its hedge has obtained. The pictures diverge with respect to what this condition is. On the first, it's some descriptive fact; on the second, it's the fact that the agent can justifiably believe that the rule in question isn't safe to rely on.

In spite of the differences between these pictures, it's possible to establish a type of correspondence between weighted contexts and a *subclass* of hedged contexts. Let's call a weighted context c and a hedged c' *equivalent* if and only if $\bigcirc X$ follows from c just in case $\bigcirc X$ follows from c'. It turns out that every weighed context c is equivalent to some hedged context, and that the hedged context that c is equivalent to has a particular sort of shape—we'll call such contexts *regular*. What's more, every hedged context having that particular shape has a corresponding weighted context that's equivalent to it. The next few pages discuss the correspondence result in more detail. In case you're not interested in these details, you can fast forward to the penultimate paragraph of this section where the significance of the result is discussed.

As a first step, we specify a procedure for deriving hedged contexts from weighted ones:

Definition 1 (Derived hedged contexts) *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be some weighted context. We con-*

succeed in showing that epistemic conflicts are not dilemmas can be stated thus: for any hedged context $c = \langle \mathcal{W}, \mathcal{R} \rangle$, for any two rules r, r' in $\mathcal{R}$ such that $contrary(r, r')$, if r and r' are both in $Triggered_c$, then either $\bigcirc Believe(Out(r))$, or $\bigcirc Believe(Out(r'))$ has to follow from c.

struct a hedged context c' from it as follows. Let c' be $\langle \mathcal{W}', \mathcal{R}' \rangle$, where

- *$\mathcal{W}' = \mathcal{W}$, and*
- *$\mathcal{R}'$ is acquired from $\langle \mathcal{R}, \leq \rangle$ by the following procedure.*

For every rule $r \in \mathcal{R}$,

1. *Let $\mathcal{R}_r = \{r' \in \mathcal{R} : r \leq r' \text{ and } contrary(r, r')\}$;*
2. *set $\mathcal{Z} = \{\neg X : \frac{X}{Y} \in \mathcal{R}_r\}$;*
3. *finally, replace $r \in \mathcal{R}$ with the hedged rule $\frac{Premise[r] : \mathcal{Z}}{Conclusion[r]}$.*

Notice that the hedged rules of the contexts obtained by this procedure contain *only* negations of the premises of rules contrary to them. Call hedged contexts of this shape *regular*.

Definition 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, for any rule r in $\mathcal{R}$, the hedge of r is the set $\{\neg Premise[r'] : r' \in \mathcal{R}'\}$ where $\mathcal{R}' \subseteq \{r' \in \mathcal{R} : contrary(r, r')\}$.*

Although *regularity* is a formal condition, it's easy to make intuitive sense of: it is a restriction on what rule hedges can do that allows them to *only* resolve clashes between rules.

When a (regular) hedged context c' is derived from the weighted context c by the procedure specified in Definition 1, c and c' are equivalent. Since this is the first part of our central result, we formulate it as an observation—its proof is given in the appendix:

Observation 6.1 *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be some weighed context and $c' = \langle \mathcal{W}', \mathcal{R}' \rangle$ a hedged context derived from it by the procedure specified in Definition 1. Then c and c' are equivalent, that is, $\bigcirc X$ follows from c just in case $\bigcirc X$ follows from c'.*

While every weighted context is equivalent to some unique regular hedged context, two weighted contexts may be equivalent to the same hedged context. As an illustration, consider the toy contexts $c_9 = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ where $\mathcal{W} = \{A, B\}$, $\mathcal{R} = \{ r_{13} = \frac{A}{C} , r_{14} = \frac{B}{D} \}$, and $r_{13} < r_{14}$, and c_{10} which is like c_9, except for its ordering on rules is reversed, that is, $r_{14} < r_{13}$. (Thus, the only difference

between c_9 and c_{10} is in the relative weights of rules.) Given that r_{13} and r_{14} aren't contrary (by assumption), the priority relation has no bearing on which ought-formulas follow from the context, and so it's natural to think of it as surplus information. It's not difficult to verify that c_9 and c_{10} are equivalent to the same hedged context $c_{11} = \langle \mathcal{W}', \mathcal{R}' \rangle$ where $\mathcal{W}' = \mathcal{W}$ and $\mathcal{R}' = \mathcal{R}$. What looked like surplus information is absent from it.

But the class of weighted contexts wouldn't be equivalent to that of regular hedged contexts, if there was some regular context without a corresponding weighted contexts. It can be shown, however, that every regular context can be derived from an equivalent weighted context: there's another procedure generating a weighted context c' from a regular hedged context c such that c' and c come out equivalent.

Definition 3 (Derived weighted contexts) *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a regular hedged context. We derive a weighted context c' from it as follows. We begin by defining an ordering on rules from $\mathcal{R}$, using their hedges: for any two rules $r, r' \in \mathcal{R}$, let*

$$r \leq r' \qquad \textit{if and only if} \qquad \neg Premise[r'] \in Hedge[r].$$

Now let c' be $\langle \mathcal{W}', \mathcal{R}', \leq \rangle$, where

- *$\mathcal{W}' = \mathcal{W}$,*
- *$\mathcal{R}' = \{ \frac{X}{Y} : \frac{X : \{\neg Z_1, \ldots, \neg Z_n\}}{Y} \in \mathcal{R} \}$, and*
- *$\leq$ is the reflexive and transitive closure of the relation between rules in $\mathcal{R}'$ defined as follows:*

  $r^* \leq r^\dagger$ if and only if $\dfrac{Premise[r^*] : \mathcal{Z}}{Conclusion[r^*]} \leq \dfrac{Premise[r^\dagger] : \mathcal{Z}'}{Conclusion[r^\dagger]}$.

The next observation expresses the second part of our central result—the proof is in the appendix:

Observation 6.2 *Let the weighted context $c' = \langle \mathcal{W}', \mathcal{R}', \leq \rangle$ be derived from some regular hedged context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ by the procedure specified in Definition 3. Then c' and c are equivalent.*

Thus, there's a many-one equivalence between weighted contexts and regular hedged ones. To see how this translates into relations between our formalized views on rules, we must start by taking a closer look at the shape of regular contexts. Recall that the hedge of any rule in such a context can contain *only* (negations of) the premises of rules that are contrary to it. Since, in our framework, rule premises always refer to descriptive features, regular contexts are the domain of our first model of the hedged-rules view. Given this, one may be tempted to take the formal result to show that our model of the contributory-rules view and the first model of the hedged-rules view are *extensionally equivalent*, or that they provide the same answer to the question of what beliefs the agent out to have in any given situation. This conclusion is on the right track, but there are two important caveats. First, I think that the result establishes more than a mere extensional equivalence. Although Observations 6.2 and 6.1 don't make it explicit, the correspondence between the two types of contexts runs fairly deep: they do not only support the same ought-formulas, but also identify the same propositions as reasons and the same conditions of defeat.[29] Second and more importantly, the result reveals a *condition* under which our first version of the hedged-rules view turns out to be (extensionally) equivalent to the contributory-rules view: the equivalence holds only in case it's assumed that any given rule's hedge can *only* refer to a subset of descriptive features, namely, those that trigger rules that are contrary to it. Thus, the model of the contributory is (extensionally) equivalent to a fragment of the descriptive version of the hedged-rules view.

Since the contributory-rules view and the descriptive hedged-rules view look very different on the surface, the result should be interesting in its own right. But it also has some significant implications, one of which we will see in the next section.

[29]Our framework is rich enough to define such notions as *epistemic reasons*, *defeat*, and *defeaters* in a mathematically precise way. I say more about this in Section 7. For an even more detailed discussion, see (Horty 2012, p. 62ff).

7 Rebutting and undercutting defeat

This section focuses on the familiar distinction between *rebutting* and *undercutting defeat*. Its main goal is to explore whether each of the three models set up in Sections 3–5 can capture it. This will let us compare their expressive power.

But let's start with what I take to be a standard characterization of the two types of defeat. Suppose some consideration X is a reason for you to believe that Y, and that another consideration Z *defeats* X. We will soon provide a formal definition of defeat. For now, you can think of it in counterfactual terms: it would be rational for you to believe that Y without Z, but it's not rational for you to believe Y with Z present. Now, Z is a *rebutting defeater* (of X as a reason to believe Y) just in case Z itself is a reason to believe not-Y. And Z is an *undercutting defeater* just in case it is *not* itself a reason to believe not-Y, but, rather, a reason to believe that X is not sufficiently indicative of the truth of Y, or, as Pollock and Cruz (1999) put it, "a reason.. to doubt or deny that Y would not be true unless X were true" (p. 196). Suppose you believe that it's raining outside on the basis of Walter's testimony, and then look outside just to see that it's raining. In this case, the fact that you perceive that it's raining rebuts Walter's testimony. As for undercutting, the go-to example is the situation where you form a believe that an object is red on the basis of perception and then learn that it is illuminated by red lights.[30]

Now let's see if this distinction can be captured in our models. And we start with the model of the descriptive version of the hedged-rules view which turns out to capture the distinction with relative ease. First off, notice that, in it, we can define defeat in a natural way. Fix a context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ and some rule $r = \dfrac{X : \neg Z_1, \cdots, \neg Z_n}{Y}$ that is triggered in c. Further, fix some subset $\mathcal{R}'$ of $\mathcal{R}$. Then we say that r is *defeated* in c, relative to $\mathcal{R}'$, just in case r is not admissible in c,

[30] Compare to Pollock's seminal discussion of the distinction in (Pollock 1974, p. 41ff) and (Pollock & Cruz 1999, pp. 196–7).

relative to $\mathcal{R}'$. (Notice that $\mathcal{R}'$ has to be fixed because our notion of admissibility is defined relative to a subset of $\mathcal{R}$. Ultimately, we're interested in which rules get defeated relative to the subset of $\mathcal{R}$ that is stable, or that determines which ought-formulas follow from the context.) Given the way we have set things up, this implies that there is a $\neg Z$ in $Hedge[r]$ such that $\mathcal{W} \cup \vdash Z$. In such a situation, we call the formula Z a *defeater* of the rule r. What's more, the model also lets us talk about reasons and defeat at their level. Here is the preliminary proposal—which will be slightly refined in a minute: whenever a rule of the form $r = \frac{X : \neg Z_1, \cdots, \neg Z_n}{Y}$ is triggered in c, we say that X is a reason for Y. In this case, we can say that X's being a reason for Y *depends on* r, or that X and Y stand in the *reason relation*. And whenever r gets defeated by some Z, we can say that X gets defeated as a reason for Y by Z, or that Z is a defeater of X as a reason for Y.[31]

Having the notion of defeat in hand, we can make use of the familiar notion of contrary rules to distinguish between rebutting and undercutting defeat as follows:[32]

- Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a hedged context, r a rule from $\mathcal{R}$ that's triggered in it, and Z a consideration that defeats this rule (relative to $\mathcal{R}' \subseteq \mathcal{R}$). Then:
  - (1) r is *rebutted* by Z in c (relative to $\mathcal{R}'$) just in case $Z = Premise[r']$ for some rule $r' \in \mathcal{R}$ such that r and r' are contrary;
  - (2) r is *undercut* by Z in c (relative to $\mathcal{R}'$) just in case there is no $r' \in \mathcal{R}$ such that $Z = Premise[r']$ and r and r' are contrary.

Notice that this definition is faithful to our informal characterization of the distinction: when a rule of the form $\frac{X : \{\neg Z_1, \ldots, \neg Z_n\}}{Believe(Y)}$ gets rebutted by some Z_i, with $1 \leq i \leq n$, there's another rule that has Z_i as a premise and $Believe(Y')$ as a conclusion, with Y and Y' inconsistent, implying that Z is a reason to believe not-Y. And when Z_i is an undercutting defeater, there's no such rule, implying that Z_i is not a reason to believe not-Y.

[31] Cf. (Horty 2012, p. 62ff).

[32] As a reminder, two rules r and r' are contrary when they support beliefs in inconsistent propositions.

Now let's turn to a concrete scenario standardly taken to involve undercutting defeat:

> **Pill**. Suppose that you have taken a pill that makes blue things look red and red things look blue. Furthermore, you have all the evidence to believe that you have taken this pill and you know how it works. Now you look at an object in front of you. It looks red.[33]

Intuitively, here you ought to believe that the object is blue. What's more, it seems that you don't have the slightest reason to believe that it's red: although, normally, an object's looking red would speaks in favor of its being red, in this case, your knowledge of the pill's workings nullifies the normal effects of something's looking red.

To see how the hedged-rules view can account for these intuitions, we express the Pill in the context $c_{12} = \langle \mathcal{W}, \mathcal{R} \rangle$. Its hard information comprises the formulas $Perceive(R)$ and $Drug$, expressing the propositions, respectively, that the object looks red to you, and that you've taken a pill, having the effects described in the passage. The context's set of rules, in turn, comprises

$$r_{15} = \frac{Perceive(R) : \neg Drug}{Believe(R)} \text{ and}$$

$$r_{16} = \frac{Perceive(R) \& Drug}{Believe(B)}.$$

The first rule is just an instance of Hedged Perception schema—which has to be present in any situation where $Perceive(R)$ obtains. The second says that you are to believe that the object is blue in case you've taken the pill and the object looks red to you. Like some of the rules we saw in Section 5, it can be thought of as a simplification compressing several rules into one.

It's easy to see that $\bigcirc Believe(B)$ follows from c_{12}, as desired. However, our preliminary approach to reasons-talk leads to the counterintuitive result that you have a reason to believe that the object is red: since r_{15} is triggered in the context, $Perceive(R)$ qualifies as a reason to believe

[33]The case is adopted from (Dancy 2017).

R, even if a defeated one. Luckily, there's a natural and simple way to avoid this result: we can require that the rule which X's being a reason for Y depends on is not only triggered, but also *not* undercut. (It may still be rebutted.) With this simple adjustment, the model delivers the intuitive result. The instance of Perception r_{15} gets defeated by the consideration $Drug$. However, given that there's no rule that's contrary to r_{15} and that has $Drug$ as a premise, r_{15} comes out undercut, entailing that $Perceive(R)$ is *not* a reason to believe that the object is red.[34] So, the descriptive version of the hedged-rules view accommodates the distinction between rebutting and undercutting defeat, as well as the Pill scenario.

But the same can't be said of the contributory-rules view. Dancy (2004, 2017) and others dismiss this view precisely on the basis of cases like the Pill.[35] What they find unacceptable is that the view entails that you have some reason to believe that the object is red, and, more generally, that certain features—such as something's looking red—*always* provide reason for having certain kinds of belief. Our formal analysis further supports this conclusion.

Notice that the correspondence result from Section 6 entails that there's no weighted context that's equivalent to the hedged c_{12}, capturing the Pill scenario: since the formula $\neg Drug$ occurs in the hedge of r_{15} and there's no rule in c_{12} that would have $Drug$ as a premise, this context is *not* regular, and only regular hedged contexts have weighted counterparts. What's more, there's a simple argument supporting the conclusion that it's impossible to capture this scenario in the contributor-rules view. Given the result, every weighted context is equivalent to some *regular* hedged context. But the only type of defeat that regular contexts allow for is rebutting defeat. And since we need undercutting defeat to capture the Pill, no weighted context can be adequate. The crucial second

[34]This adjustment may come across as ad hoc here, but the notion of a reason that it gives us is better aligned with the literature than the preliminary one.

[35]Such cases play a crucial role in Dancy's overall argument against generalism, or, roughly, the view according to which there are at least some (epistemic, moral) principles. See also (Bradley 2019, pp. 8–9).

premise is supported by the following observation:

Observation 7.1 *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, relative to some $\mathcal{R}' \subseteq \mathcal{R}$, it is rebutted.*[36]

A proponent of the contributory-rules view might try to resist the argument by contending that the unfortunate conclusion follows because the model doesn't capture the view in full. I doubt that this line of argument can succeed, however, and for two reasons. First, it's hard to see what's missing from the model. And second, the one author who is naturally seen as having pursued just this line arrives at a complex view that's most naturally understood as a *mixed view* on rules on which rules are contributory and can also have hedges.[37] Such a view is interesting and worth exploring, but it doesn't really help the *simple* contributory-rules view that we started with.[38]

My overall conclusion with respect to the model of the contributory-rules view, then, is this:

[36] Proof sketch: Suppose r is defeated in c, relative to some $\mathcal{R}' \subseteq \mathcal{R}$. This means that there's some Z such that $\neg Z \in Hedge[r]$ and $\mathcal{W} \subseteq Conclusion[\mathcal{R}'] \vdash Z$. Since c is regular, we know that $Hedge[r] \subseteq \{\neg Premise[r'] : r' \in \mathcal{R} \text{ and } contrary(r, r')\}$. 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.

[37] The author is Horty (2012). To get his model to deliver the intuitive result in the Pill, he appeals to a special type of *exclusionary rules* that take other rules out of consideration—see, especially, pp. 231–2. But any situation where a rule r gets taken out of consideration by an exclusionary rule r' can be thought of as a situation where r is a hedged rule, with $\neg Believe(Out(r))$ listed in its hedge, and the conclusion of r' is $Believe(Out(r))$. Thus, Horty's exclusionary rules seem to be closely connected to the hedged rules we discussed in Section 5. I plan to explore these connections in future work.

[38] Another way for the proponents of the view to block the unwelcome result is to modify the conditions under which X qualifies as a reason for Y, requiring that the rule that X's being a reason for Y is not defeated at all. But while this does secure the correct result in the Pill scenario, it also looks completely ad hoc. What's more, it deprives the contributory-rules view of one of its main selling points in ethics, namely, its ability to account for *residual reasons*, or, roughly, reasons that have some power to remain even if defeated. In fact, the notion of a reason this move commits one to is so revisionary that it's hard to recognize what are standardly thought of as reasons in it.

it doesn't have the resources to distinguish between rebutting and undercutting defeat, and it has very little hope to account for such scenarios as the Pill.

Turning to the normative version of the hedged-rules view, we start by recalling that the Perception schema was expressed in it as follows:

$$r'(X) = \frac{Perceive(X) : \neg Believe(Out(r'))}{Believe(X)}.$$

Given this, it seems natural to capture the Pill scenario in the context $c_{13} = \langle \mathcal{W}, \mathcal{R} \rangle$, where $\mathcal{W}$ is, again, comprised of $Perceive(R)$ and $Drug$, and where $\mathcal{R}$ includes the following three rules:

$$r_{17} = \frac{Perceive(R) : \neg Believe(Out(r_{17}))}{Believe(R)},$$

$$r_{18} = \frac{Drug : \neg Believe(Out(r_{18}))}{Believe(Out(r_{17}))},$$

$$r_{19} = \frac{Perceive(R) \& Drug : \neg Believe(Out(r_{19}))}{Believe(B)}.$$

It's not difficult to verify that $\bigcirc Believe(B)$ follows from c_{13}: the only set of rules that's stable in it is $\{r_{18}, r_{19}\}$ and $Conclusion[\{r_{18}, r_{19}\}] = \{Believe(B), Believe(Out(r_{17}))\}$. What's more, $Perceive(R)$ doesn't qualify as a reason for $Believe(R)$ in c_{13}, since the underlying rule r_{17} gets undercut: the formula $Believe(Out(r_{17}))$ is a consequence of $\mathcal{W} \cup Conclusion[\{r_{18}, r_{19}\}]$, and there's no rule that has it as a premise.

So, the model captures the Pill scenario with ease, and, more generally, it seems to be well suited to account for cases involving undercutting defeat. But can it also distinguish them from cases involving rebutting defeat? I on my part can't think of a natural way for it do so. For starters, given the form it assigns to epistemic rules and our definition of rebutting defeat, a rule r could be rebutted by another one just in case this other rule had $Believe(\neg Out(r))$ as both premise and conclusion, and it just doesn't seem plausible that such a rule would instantiate any sensible epistemic rule schema. More importantly—and independently of our definition of rebutting—we shouldn't forget what the view captured by the model itself says. On it, whenever two epistemic

rules (seem to) come into conflict, the agent is justified to believe that at least one of them is not safe to rely on. But a rule's not being safe to rely on looks very similar to the informal characterization of undercutting defeat we saw at the outset of this section: to me, being justified to believe that a rule schema which would normally make it rational to believe that Y on the basis of X is not safe to rely on sounds much like having a (conclusive) reason to "doubt or deny that Y would not be true unless X were true" (Pollock & Cruz 1999, p. 196). So I think that it's in the nature of the view under discussion to construe any given case of defeat as undercutting. And the fact that, in our model of the view, any rule r has to proceed by way of the formula $\neg Believe(Out(r))$ listed in its hedge just makes that very explicit.

We could certainly find a place for rebutting defeat by extending the model. Thus, we could allow for the occurrence of formulas of other forms in rule hedges, or, perhaps, supplement the model with a priority relation over rules. Such extensions would be interesting and worthwhile to explore. However, my take on it is that it would also be a departure from the view we started with in the direction of mixed views on epistemic rules.

8 Concluding remarks

My main focus in this paper was on developing models of views on epistemic rules that come with two different responses to the problem of epistemic conflicts. I set up three models: the first is meant to capture a view on which rules are contributory; the second a view on which rules are hedged, with their hedges listing descriptive features which, if instantiated, make the rule inapplicable; and the third model is meant to capture a view on which a rule fails to apply when an agent can justifiably believe that it's not safe to rely on. Once the models were set up, we saw that there's a clear sense in which the first model is equivalent to a fragment of the second one. Finally, we looked at how each model fairs with respect to the familiar distinction between

two types of defeat—rebutting and undercutting—and I concluded that only the second model can accommodate the distinction, while the other two allow for one type of defeat only.

Now let's redirect attention from the three models of views on rules to the views themselves. Assuming that nothing of substance is lost due to the simplifying assumptions built into the models, and that the models represent the views adequately, our formal exploration supports the following conclusions. First, while the descriptive version of the hedged-rules view resolves the problem of epistemic conflicts and accommodates the distinction between the two types of defeat, it also appears to reduce to a version of epistemic particularism. Thus, the version of Testimony that this view ends up with and that is finitely statable runs thus: "If an agent's epistemic situation includes testimony that X, then the agent ought to believe that X—unless something comes in the way". This, however, looks like a generalization, as opposed to a genuine principle of the sort that generalists are after. Second, in spite of the surface differences, there's a clear sense in which the contributory-rules view corresponds to a restricted form of the hedged-rules view: the two are extensionally equivalent, and they identify the same considerations as reasons and defeaters. The restriction in the hedged-rules view has to do with the sorts of descriptive features that can occur in rule hedges: they have to be the features that serve as premises of other rules. Third, the contributory-rules view is committed to treating any defeat as rebutting, and, thus, it cannot accommodate the distinction between rebutting and undercutting defeat. (This follows from the correspondence. What doesn't follow are any strong conclusions about the contributory-rules view's relations to particularism, since it corresponds to a *fragment* of the apparently particularist hedged-rules view.) Fourth, the most promising version of the hedged-rules view appears to be committed to treating any defeat as undercutting, and so it too cannot accommodate the distinction between rebutting and undercutting.

Where does this leave us? Well, assuming that a satisfactory account of epistemic rules must

not be committed to particularism and also has to accommodate the distinction between rebutting and undercutting defeat, we'd seem to be out of luck. For none of the three accounts we have explored satisfy both desiderata. That being said, there is also a clear path forward: we may want to explore a mixed view on epistemic rules that would combine the idea that rules are contributory and the core idea behind the normative version of the hedged-rules view. Such a view promises to be able to accommodate the distinction between rebutting and undercutting defeat with ease; and given its simple take on the contents of rule hedges, it doesn't face the complexity problems that plague the descriptive version of the hedged-rules view.[39] So, for those of us who want to hold onto the idea that rules like Perception and Testimony are genuine, this mixed view is the best bet.

I see several directions for future research. First off, it would pay to develop and explore a model of this mixed view. We should also relate it to the models set up here, as well as kindred views explored by other authors.[40] Second, it may be worthwhile to generalize our results to richer contexts, bolstering the conclusions drawn in the last paragraphs. A third promising direction of research is to explore other applications of the framework set up here: in particular, I think it should be applied to the debate about the shape of moral principles. Relatedly, it may pay to explore further applications of our central result. Thus, according to received wisdom, contributory moral principles can, while hedged ones cannot account for cases involving residual reasons—or, roughly, reasons that have some power to remain even if defeated—and cases involving composing reasons—or situations where two (or more) weaker considerations combine to jointly defeat a consideration that's stronger than each of them taken in isolation.[41] Assuming our correspondence

[39] One can still worry that this view might end up being particularist: thus, it might end up having to postulate too many rules when accommodating various epistemic situations. We can't address this worry here, leaving it as one of the important tasks for future work.

[40] Echoing footnotes 7 and 37, I think that both Horty and Pollock have developed versions of the mixed view, even if it's not immediately apparent.

[41] See, e.g., (Dancy 2004, pp. 22–9).

result extends to views on moral principles, it would suggest that this common wisdom can't be right. So it'd be interesting to explore the issue in detail.

Appendix: Proofs of central observations

To cut the clutter in proofs, we begin by introducing the notion of a rule's *counterpart* in a different context:

- Let r be a rule of the form $\frac{X}{Y}$ and $c = \langle \mathcal{W}, \mathcal{R} \rangle$ some hedged 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 (hedged) *counterpart* of r in the context c, written as $counterpart_c(r) = r'$.

  Let r be a hedged rule of the form $\frac{X : \mathcal{Z}}{Y}$ 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 $counterpart_c(r) = r'$.

Next, we prove a lemma:

Lemma *Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a regular hedged context. Then for any sets of rules $\mathcal{R}'$ and $\mathcal{R}''$ that are stable in c, we have $\mathcal{R}' = \mathcal{R}''$.*

Proof. Suppose toward a contradiction that $\mathcal{R}' \neq \mathcal{R}''$. Then there is some rule r such that $r \in \mathcal{R}''$ and $\mathcal{R} \notin \mathcal{R}'$. Since $r \notin \mathcal{R}'$ and $\mathcal{R}'$ is stable, we can be sure that $\mathcal{R}' \notin Admissible_c(\mathcal{R}')$. Since $r \in \mathcal{R}''$ and $\mathcal{R}''$ is stable, we know $r \in Triggered_c$. Given that $\mathcal{R}' \notin Admissible_c(\mathcal{R}')$ and $r \in Triggered_c$, there has to be some $\neg Z \in Hedge[r]$ such that $\mathcal{W} \cup Conclusion[\mathcal{R}'] \vdash Z$. Since c is a regular context, $Z = Premise[r']$ where r' is some rule from such that $contrary(r, r')$. Further, given that $Conclusion[\mathcal{R}']$ can contain only $Believe$-formulas and $Premise[r']$ is not a $Believe$-formula, we have can be sure that $\mathcal{W} \vdash Premise[r']$. By the monotonicity of classical logic, $\mathcal{W} \cup Conclusion[\mathcal{R}''] \vdash Premise[r']$. Given this and the fact that $Premise[r'] \in Hedge[r]$, we can conclude that r is not admissible in $\mathcal{R}''$, and, thus, that $\mathcal{R}''$ is not stable in c after all. QED

With the notion of counterparts and the above lemma in hand, we can turn to the proofs of our two observations.

Observation 6.1 *Let $c = \langle \mathcal{W}, \mathcal{R}, \leq \rangle$ be some weighed context and $c' = \langle \mathcal{W}', \mathcal{R}' \rangle$ a hedged context derived from it by the procedure specified in Definition 1. Then c and c' are equivalent, that is, $\bigcirc X$ follows from c just in case $\bigcirc X$ follows from c'.*

Proof.

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 $r \in Binding_c$, which, in turn, entails that $\mathcal{W} \vdash Premise[r]$ and that there's no $r' \in \mathcal{R}$ such that $r' \in Triggered_c$, $contrary(r, r')$, and $r \leq r'$. By Definition 1, we can be sure that there's a rule $r' \in \mathcal{R}'$ such that $counterpart_c(r') = r$. Our Lemma tells us that there's a unique set of rules $\mathcal{R}''$ that is stable in c'. If we can show that $r' \in \mathcal{R}''$, then we are done. So suppose that it isn't. Since $\mathcal{W} \vdash Premise[r]$ and $\mathcal{W}' = \mathcal{W}$, we can be sure that $r' \in Triggered_{c'}$. So, r' is not admissible in $\mathcal{R}''$ because there's a $\neg Z \in Hedge[r']$ such that $\mathcal{W} \cup Conclusion[\mathcal{R}''] \vdash Z$. By Definition 1, there's a rule r^* such that $Z = Premise[r^*]$, $contrary(r, r^*)$, and $r \leq r^*$. Since $Conclusion[\mathcal{R}'']$ contains only $Believe$-formulas and $Premise[r^*]$ is not a $Believe$-formula, we can be sure that $\mathcal{W}' \vdash Premise[r^*]$. And given that $\mathcal{W} = \mathcal{W}'$, we have $\mathcal{W} \vdash Premise[r^*]$. But from $\mathcal{W} \vdash Premise[r^*]$ it follows that $r^* \in Triggered_c$. This together with the facts that $contrar(r, r^*)$ and $r \leq r^*$ is enough to conclude that $r \notin Binding_c$, 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 $r \in \mathcal{R}''$ where $\mathcal{R}''$ is the unique set of rules that is stable in c'. (We know that $\mathcal{R}''$ is unique by Lemma 1.) Since $\mathcal{R}''$ is stable, we know $r \in Admissible_{c'}(\mathcal{R}'')$, that is, that $r \in Triggered_{c'}$ and that there's is no $\neg Z \in Hedge[r]$ such that $\mathcal{W}' \cup Conclusion[\mathcal{R}''] \vdash Z$. By Definition 1, we know that there's a rule $r' \in \mathcal{R}$ such that $counterpart_c(r) = r'$. Notice that if we can show that $r' \in Binding_c$, we will be done. So suppose the opposite: that $r' \notin Bindind_c$.

Given that $Premise[r] = Premise[r']$, $\mathcal{W}' = \mathcal{W}$, and $\mathcal{W}' \vdash Premise[r]$, it has to be the case that $r' \in Triggered_c$. So r' is not in $Binding_c$ because there is another rules $r'' \in \mathcal{R}$ such that $r'' \in Triggered_c$, $contrary(r', r'')$, and $r' \leq r''$. By Definition 1, we can be sure that $\neg Premise[r''] \in Hedge[r]$. And since $\mathcal{W} \vdash Premise[r'']$ and $\mathcal{W}' = \mathcal{W}$, we know that $\mathcal{W}' \vdash Premise[r'']$. By the monotonicity of classical logic, we also have $\mathcal{W}' \cup Conclusion[\mathcal{R}''] \vdash Premise[r'']$. Thus, after all, the hedge of r does contain a formula of the form $\neg Z$ (namely, $\neg Premise[r'']$) for which we have $\mathcal{W}' \cup Conclusion[\mathcal{R}''] \vdash Z$. QED

Observation 6.2 *Let the weighted context $c' = \langle \mathcal{W}', \mathcal{R}', \leq \rangle$ be derived from some regular hedged context $c = \langle \mathcal{W}, \mathcal{R} \rangle$ by the procedure specified in Definition 3. Then c' and c are equivalent.*

Proof.

Let $c'' = \langle \mathcal{W}'', \mathcal{R}'' \rangle$ be the result of applying Definition 1 to c'. By Observation 6.1, the contexts c' and c'' are equivalent. What we need to do is show that $c = c''$.

(1) It's obvious that $\mathcal{W} = \mathcal{W}''$.

(2) $\mathcal{R} \subseteq \mathcal{R}''$: Consider an arbitrary rule r from $\mathcal{R}$. By Definition 3, there has to be some rule r' in $\mathcal{R}'$ such that $counterpart_c(r') = r$. Definition 1 tells us that there is a rule r'' in $\mathcal{R}''$ such that $counterpart_{c'}(r'') = r'$. It's clear that $Premise[r''] = Premise[r]$ and $Conclusion[r''] = Conclusion[r]$. So we only need to show that $Hedge[r''] = Hedge[r]$:

- $Hedge[r] \subseteq Hedge[r'']$: Consider an arbitrary $Z \in Hedge[r]$. Since c is regular (by assumption), $Z = \neg Premise[r^*]$ for some rule $r^* \in \mathcal{R}$ such that $contrary(r, r^*)$. By Definition 3, there has to be a rule $r^\dagger \in \mathcal{R}'$ such that $counterpart_c(r^\dagger) = r^*$ and $r' \leq r^\dagger$. Notice that $contrary(r', r^\dagger)$. Now, by Definition 1, $r^\dagger$ is in $\mathcal{R}_{r'}$, and so $\neg Premise[r^\dagger] \in Hedge[counterpart_{c''}(r')]$. But $counterpart_{c''}(r') = r''$ and $Premise[r^\dagger] = Premise[r^*]$ together imply that $Z \in Hedge[r'']$.

- $Hedge[r''] \subseteq Hedge[r]$: Consider some $Z \in Hedge[r'']$. By Definition 1, $Z = \neg Premise[r^*]$ for some rule $r^* \in \mathcal{R}'$ such that $r^* \in \mathcal{R}'_{r'}$ and $counterpart_{c'}(r'') = r'$. Since $r^* \in \mathcal{R}'_{r'}$, we know that $r' \preceq r^*$ and $contrary(r', r^*)$. By Definition 3, we can be sure that $counterpart_c(r') = r \preceq counterpart_c(r^*)$, and, from $r \preceq counterpart_c(r^*)$, that $\neg Premise[counterpart_c(r^*)]$ is in $Hedge[r]$. But $Premise[counterpart_c(r^*)] = Premise[r^*]$, and so $Z \in Hedge[r]$.

(2) $\mathcal{R} \supseteq \mathcal{R}''$: Similar to the other direction.

QED

References

Boghossian, P. A. (2008). Epistemic rules. *Journal of Philosophy*, 105(9), 472–500.

Bradley, D. (2019). Are there indefeasible epistemic rules? *Philosopher's Imprint*, 19(3), 1–19.

Chisholm, R. (1980). A version of foundationalism. *Midwest Studies in Philosophy*, 5(1), 543–64.

Christensen, D. (2007). Does Murphy's Law apply in epistemology? Self-doubt and rational ideals. *Oxford Studies in Epistemology*, 2, 3–31.

Christensen, D. (2010). Higher-order evidence. *Philosophy and Phenomenological Research*, 81(1), 185–215.

Christensen, D. (2013). Epistemic modesty defended. In D. Christensen & J. Lackey (Eds.), *The Epistemology of Disagreement: New Essays* (pp. 77–97). Oxford University Press.

Conee, E. & Feldman, R. (2004). *Evidentialism: Essays in Epistemology*. 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.

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.

Engel, P. (2013). In defense of normativism about the aim of belief. In T. Chan (Ed.), *The Aim of Belief* (pp. 32–63). Oxford University Press.

Goldman, A. (1986). *Epistemology and Cognition*. Harvard University Press.

Goldman, A. (2009). Internalism, externalism, and the architecture of justification. *The Journal of Philosophy*, 106(6), 309–38.

Holton, R. (2002). Principles and particularisms. In *Proceedings of the Aristotelian Society, Suppl. Volume 76* (pp. 191–210).

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.

Lasonen-Aarnio, M. (2010). Unreasonable knowledge. *Philosophical Perspectives*, 24, 1–21.

Lasonen-Aarnio, M. (2014). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, 88(2), 314–45.

Makinson, D. (2005). *Bridges from Classical to Nonmonotonic Logic*. King's College Publications.

Peacocke, C. (2004). *The Realm of Reason*. Oxford University Press.

Pollock, J. (1974). *Knowledge and Justification*. Princeton: Princeton University Press.

Pollock, J. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA: MIT Press.

Pollock, J. & Cruz, J. (1999). *Contemporary Theories of Knowledge*. Rowman & Littlefield Publishers.

Pryor, J. (2000). The skeptic and the dogmatist. *Noûs*, 34, 517–49.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligece*, 13, 81–132.

Ross, W. D. (1930). *The Right and the Good*. Oxford University Press.

Scanlon, T. M. (2000). Principles and particularisms. In *Proceedings of the Aristotelian Society, Suppl. Volume 74* (pp. 301–17).

Titelbaum, M. (2015). Rationality's fixed point (or: in defense of right reason). *Oxford Studies in Epistemology*, 5, 253–294.

Wedgwood, R. (2002). Internalism explained. *Philosophy and Phenomenological Research*, 65, 349–69.