Forthcoming in *Synthese.*
Please cite the published version.

Conciliatory views, higher-order disagreements, and defeasible logic

Aleks Knoks

Abstract Conciliatory views of disagreement say, roughly, that it's rational for you to become less confident in your take on an issue in case you find out that an epistemic peer's take on it is the opposite. Their intuitive appeal notwithstanding, there are well-known worries about the behavior of conciliatory views in scenarios involving higher-order disagreements, which include disagreements over these views themselves and disagreements over the peer status of alleged epistemic peers. This paper does two things. First, it explains how the core idea behind conciliatory views can be expressed in a defeasible logic framework. The result is a formal model that's particularly useful for thinking about the behavior of conciliatory views in cases involving higher-order disagreements. And second, the paper uses this model to resolve three paradoxes associated with disagreements over epistemic peerhood.

Keywords Disagreement· Epistemic peer· Conciliationism· Self-undermining· Defeasible logic

1 Introduction

Think of your favorite philosophical problem. You've likely pondered on it for a long time, and you must have views on how to solve it. And odds are you know someone who has pondered on it equally long, whose credentials are as good as yours, and whose views directly oppose yours. If so, you're in disagreement with an *epistemic peer*.[1] But what should you make of this disagreement? And what effects

Aleks Knoks, https://orcid.org/0000-0001-8384-0328
University of Luxembourg
E-mail: aleks.knoks@uni.lu

[1] Although there are a number of accounts of epistemic peerhood, central to all of them is the idea that an epistemic peer is an epistemic equal. Since (Kelly, 2005), it's standard to emphasize two kinds of equality: the equality of evidence and the equality of the peer's capacity to process this evidence. Epistemic peerhood is also a relative notion: The fact that two people are epistemic equals in matters pertaining to metaphysics doesn't mean that they are equals in matters pertaining to contemporary politics. I will rely on

should it have on your take on the problem? Advocates of *conciliatory views* are pulled to the idea that this disagreement is epistemically significant for you, and that it should make you less confident that your take on the problem is correct. Advocates of *steadfast views*, by contrast, are pulled to the idea that this disagreement might not be that significant, and that there might be nothing wrong with you remaining fully confident in the correctness of your views.[2]

The conciliatory views are more intuitive: Given that the problem is complex, that you are fallible, and that the most straightforward explanation for the existence of the disagreement is that either your views, or those of your opponent rest on a subtle mistake, it seems only reasonable for you to lose some of your confidence.[3] But their intuitive appeal notwithstanding, there are also serious worries about the behavior of conciliatory views in scenarios involving what we might call *higher-order disagreements*. Suppose that you're a committed conciliationist and that you find yourself in disagreement over X with an epistemic peer, Francis. Now, how should you respond to your disagreement with Francis, in case you find yourself in a further disagreement over the truth of conciliationism? And what should you do, in case there's another peer who thinks that Francis is not your epistemic peer?[4] Adam Elga (2010) has forcefully argued that, when trying to respond to the first question, conciliatory views turn on themselves and, thereby, issue inconsistent directives: On the one hand, they seem to say that you should lose confidence in X. On the other, they seem to say that you should lose confidence in conciliationism itself, and, thereby, indirectly that you should *not* lose confidence in X.[5] And more recently, Thomas Mulligan (2015) has argued that conciliatory views produce "odd, incoherent, or downright paradoxical results" when trying to respond to the second question too: Thus, on the one hand, they seem to say that you should lose confidence in Francis' status as a peer. On the other, they also seem to imply that the disagreement over Francis' status serves as

an intuitive understanding of the notion throughout this paper. For more on it, see (Gelfert, 2011), (King, 2012), (Matheson, 2015b, Ch. 2), and (Matheson, 2018).

[2] Well-known advocates of conciliatory views include Christensen (2007, 2011, 2016), Elga (2007), Feldman (2005, 2006, 2009), and Matheson (2015b); those of steadfast views include Kelly (2005, 2010) and Titelbaum (2015).

[3] Since your opponent is your epistemic equal, it's as likely that you've made a mistake as it is that she has made one. So there's no reason for you to think that your views are more likely to be correct.

[4] Why call these disagreements *higher-order*? Notice that the first scenario involves a disagreement about whether or not disagreements have any epistemic significance: You think they do, your peer thinks they don't. And the second scenario involves a disagreement about whether or not a particular disagreement you find yourself in is epistemically significant for you: You think it is, your peer thinks it is not.

[5] I'm glancing over details here. The literature talks about at least four distinct (albeit related) worries associated with cases involving disagreements over conciliatory views: First, there's the worry that a conciliationist must abandon her view by her own lights—see (Decker, 2014; Kelly, 2005; Littlejohn, 2013; Matheson, 2015a,b). Second, there's the related worry that conciliationism issues inconsistent directives—see (Christensen, 2013; Decker, 2014; Elga, 2010; Littlejohn, 2013, 2020; Matheson, 2015a,b; Weatherson, 2013). Third, there's the worry that a conciliationist has to abandon her view when repeatedly disagreeing about conciliationism with a stubborn opponent—see (Decker, 2014; Elga, 2010; Weatherson, 2013)—and, fourth, the worry that a conciliationist can't maintain a stable view on how to respond to disagreement—see (Christensen, 2013; Weatherson, 2013). The first two worries are the most pressing ones. For attempts to respond to them, see (Bogardus, 2009, pp. 332–3), (Christensen, 2013, pp. 90–6), (Christensen, 2021), (Elga, 2010, Secs. 7–8), (Fleisher, 2021), (Knoks, forthcoming), (Littlejohn, 2020), (Matheson, 2015a), (Matheson, 2015b, pp. 153–7), and (Pittard, 2015).

good evidence that you've overestimated your ability to assess peerhood, and that, in response to it, you should lose your confidence in this ability. The trouble is that your judgment about the existence of the disagreement over Francis' status itself relies on this ability, and that there's no ground to lose confidence in Francis' status as a peer, once you've lost confidence in this ability. Thus, conciliatory views would seem to imply both that you should lose confidence in Francis' status and that you should not.

Of course, this is sketchy and quick, but you should agree that advocates of conciliationism can't ignore scenarios involving higher-order disagreements, and that there's something here that they need to explain. I've two goals in this paper. The first is to show how the core idea behind conciliatory views can be expressed in a defeasible logic framework, resulting in a formal model which is particularly useful for thinking about the behavior of conciliatory views in cases involving higher-order disagreements.[6] The second goal is to use this model to respond to Mulligan's worries, which—unlike the problems that Elga identified—seem to have gone unnoticed in the literature. More specifically, we'll aim to resolve the three paradoxes for conciliatory views having to do with disagreements over peerhood that Mulligan has identified—only one of which was sketched in the previous paragraph.

The remainder of this paper is structured as follows. Section 2 presents a very simple general-purpose *defeasible reasoner*, or a (defeasible) logic with a consequence relation at its core. Section 3 embeds the core of conciliationism in it. Section 4 then looks at the puzzling case Mulligan uses to elicit two of the paradoxes in more detail, explains how it can be expressed in the model, and, thereby, also resolves the two paradoxes. Section 5 refocuses on a different case (which Mulligan uses to elicit the third paradox), shows that the phenomenon that makes this case puzzling isn't specific to conciliationism, and argues that it isn't nearly as problematic as one may think. Section 6 concludes.

2 Basic defeasible reasoner

The defeasible reasoner we define in this section is a form of *default logic*.[7] The core idea behind default logic is to supplement classical logic with a set of special *default rules*, or simply *defaults*, representing defeasible generalizations, so that a stronger set of conclusions can be obtained from a given set of premises.

As background, we assume the language of ordinary propositional logic with all the usual connectives, and we let the customary turnstile symbol $\vdash$ stand for classical logical consequence. Default rules are represented as pairs of (vertically) ordered formulas: Where X and Y are arbitrary propositions, $\frac{X}{Y}$ stands for the rule that lets us conclude Y from X by default. To take an example, let LB and B stand for the propositions, respectively, that the object you're looking at looks blue to you, and that it is blue. Then $\frac{LB}{B}$ lets us conclude that the object is blue on the basis of its

[6] More specifically, we model a particularly strong conciliatory view, coming close to the *Equal Weights View*, according to which, in a case of disagreement over X, you are to give a peer's confidence in X the same weight as your own—see, e.g., (Elga, 2007), (Matheson, 2015b).

[7] The original formulation of default logic is due to Reiter (1980). My presentation draws on the more user-friendly version of Horty (2012).

looking blue to you. Notice that this rule is naturally thought of as an instance of the sensible, yet defeasible, principle that things that look blue usually are blue.

In addition to vanilla defaults, we're going to have special *exclusionary defaults* that can exclude other rules, or take them out of consideration. In order to formulate such rules, we extend the background language in two ways. First, we introduce rule names, assigning every default rule a unique name—and we use the letter r here, with subscripts. Second, we introduce a predicate $Out(\cdot)$ that takes rule names as arguments. The intended meaning of the formula $Out(r)$ is that the rule that r refers to is excluded or taken out of consideration. Thus, if we let r_1 be the name of the above rule $\frac{LB}{B}$, the formula $Out(r_1)$ says that r_1 is excluded.[8]

In order to be in a position to pick out the premises and conclusions of default rules, we introduce the functions $Premise[\cdot]$ and $Conclusion[\cdot]$. So where $r = \frac{X}{Y}$ is some rule, $Premise[r] = X$ and $Conclusion[r] = Y$. The second function will sometimes get applied to sets of rules too: Where $\mathcal{S}$ is a set of defaults, $Conclusion[\mathcal{S}]$ is the collection of conclusions of all rules in $\mathcal{S}$, that is, $Conclusion[\mathcal{S}] = \{Conclusion[r] : r \in \mathcal{S}\}$.

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 the agent is certain of—and $\mathcal{R}$ is a set of default rules—the rules the agent relies on when reasoning. We call such structures *contexts* and denote them by the letter c, with subscripts. Our first two contexts will capture the following toy scenario—call it *Blue Lights*: You're looking at an object in front of you. It looks blue. Then you learn that the object is illuminated by blue lights. Since the scenario unfolds in two steps, we represent it using two contexts $c_1 = \langle \mathcal{W}, \mathcal{R} \rangle$ and $c_2 = \langle \mathcal{W}', \mathcal{R} \rangle$. Let LB and B be as before, and let I say that the object is illuminated by blue lights. The hard information of the first context $\mathcal{W}$ is comprised of LB, while that of the second $\mathcal{W}'$ is comprised of LB and I. The set of rules $\mathcal{R}$ is the same throughout, and it is comprised of the familiar rule $r_1 = \frac{LB}{B}$ and the new exclusionary rule $r_2 = \frac{I}{Out(r_1)}$. Intuitively, the latter says that r_1 can't be relied on and is to be excluded, if it turns out that the object is illuminated by blue lights.

We will now specify a procedure that will determine which formulas follow from any given context c. Where $c = \langle \mathcal{W}, \mathcal{R} \rangle$ is a context, let any subset $\mathcal{S}$ of $\mathcal{R}$, $\mathcal{S} \subseteq \mathcal{R}$, be called a *scenario based on* c. Our procedure will rely on an intermediary notion of a *proper scenario*. Intuitively, a proper scenario based on a context $\langle \mathcal{W}, \mathcal{R} \rangle$ is a special subset of $\mathcal{R}$ that includes all and only those rules from $\mathcal{R}$ which should apply or should be in force in the case c represents. The actual definition will emerge as a combination of two other concepts. The first formalizes the intuitive idea that a rule should be in force only if its premise is derivable:

- 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 $Triggered_{\mathcal{W},\mathcal{R}}(\mathcal{S}) = \{r \in \mathcal{R} : \mathcal{W} \cup Conclusion[S] \vdash Premise[r]\}$.

[8] For simplicity, I use the same names to refer to rules in our extended propositional language as I do when talking about them in English. Compare to (Horty, 2012, Sec. 5.2) where different names are used.

Applying this definition to the empty scenario $\varnothing$, against the background of the context c_1, we get $Triggered_{\mathcal{W},\mathcal{R}}(\varnothing) = \{r_1\}$: Since the formula $LB = Premise[r_1]$ follows from the context's hard information $\mathcal{W} = \{LB\}$, it follows from the set $\mathcal{W} \cup Conclusion[\varnothing]$ too. And, thus, r_1 qualifies as triggered in $\varnothing$. The formula $I = Premise[r_2]$, on the other hand, doesn't follow from $\mathcal{W} \cup Conclusion[\varnothing]$, and, thus, r_2 doesn't qualify as triggered in $\varnothing$. (You may wonder why we opted for $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Premise[r]$ in the definition, as opposed to the simpler $\mathcal{W} \vdash Premise[r]$. The answer is that the more complex expression is needed to handle situations where some rules get triggered not by the context's hard information, but by other rules that are triggered themselves. We'll see many examples of this in the next section.)

Now let's consider the context c_2, representing the second episode of Blue Lights. It can be verified that both $r_1 = \dfrac{LB}{B}$ and $r_2 = \dfrac{I}{Out(r_1)}$ qualify as triggered in every scenario based on it. But, intuitively, r_1 shouldn't be in force. Intuitively, r_1 should be excluded by r_2. This simple idea motivates the second concept our definition of proper scenarios will rely on:

- 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 scenario $\mathcal{S}$ are those that belong to the set $Excluded_{\mathcal{W},\mathcal{R}}(\mathcal{S}) = \{r \in \mathcal{R} : \mathcal{W} \cup Conclusion[\mathcal{S}] \vdash Out(r)\}$.

Applying this concept to the scenario $\{r_1, r_2\}$, against the background of c_2, we get $Excluded_{\mathcal{W}',\mathcal{R}}(\{r_1, r_2\}) = \{r_1\}$: Since $\mathcal{W}' \cup Conclusions[\{r_1, r_2\}]$ is the set $\{LB, I, B, Out(r_1)\}$, the formula $Out(r_1)$ is among its (classical) consequences, while the formula $Out(r_2)$ is not. So r_1 does, while r_2 doesn't qualify as excluded in $\{r_1, r_2\}$.

With the concepts of triggered and excluded rules in hand, we have what we need to define proper scenarios:[9]

- Let $\langle \mathcal{W}, \mathcal{R} \rangle$ be a context and $\mathcal{S}$ a set of default rules based on this context. Then $\mathcal{S}$ is a *proper scenario* based on $\langle \mathcal{W}, \mathcal{R} \rangle$ just in case

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

According to this definition, a proper scenario $\mathcal{S}$ contains all and only those rules that are triggered *and* not excluded in it. Let's verify that it gives us the intuitively correct result for c_1 and c_2: Since these two contexts share their set of rules, the same sets qualify as scenarios based on them, namely, $\varnothing$, $\{r_1\}$, $\{r_2\}$ and $\{r_1, r_2\}$. Now,

[9] Two notes about this definition: First, I chose to keep it as simple as possible, while having the goals of this paper in mind. So one shouldn't be surprised that it isn't well-equipped to handle arbitrary scenarios. Note, however, that nothing stands in the way of refining it further. In particular, we could incorporate Horty's (2012) concepts of *conflicting* and *defeated rules* into it, which would let it handle cases where two triggered rules let us derive inconsistent propositions, as well as cases where one of such two rules appears to have more force than the other—see (Horty, 2012, pp. 40–5). Second, the definition ignores a technical problem that has to do with aberrant contexts containing self-triggering chains of rules, such as $\langle \mathcal{W}, \mathcal{R} \rangle$ with $\mathcal{W} = \varnothing$ and $\mathcal{R} = \{\dfrac{A}{A}\}$. Nothing important hinges on this, however. See (Horty, 2012, p. 48f & Appendix A.1) for a discussion of this problem and a solution to it.

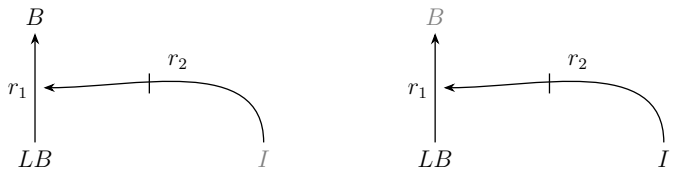

Fig. 1 Blue Lights, c_1 (left) and c_2 (right)

$\varnothing$ doesn't qualify as a proper scenario based on c_1 because it fails to contain the rule r_1 that's triggered and not excluded in it; and $\{r_2\}$ and $\{r_1, r_2\}$ don't qualify because both contain the rule r_2 that's not triggered in them. This leaves $\{r_1\}$ as the unique proper scenario based on c_1. As for c_2, here $\varnothing$ and $\{r_1\}$ don't qualify because both fail to contain a rule that's triggered and not excluded in them, namely, r_2; and $\{r_1, r_2\}$ doesn't qualify because it contains a rule that's excluded in it, namely, r_1. This leaves $\{r_2\}$ as the unique proper scenario based on c_2.

Our final definition specifies which formulas follow from a context:

- Let $c = \langle \mathcal{W}, \mathcal{R} \rangle$ be a context. Then the statement X follows from c just in case $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$ for *each* proper scenario $\mathcal{S}$ based on c.[10]

Notice that this definition delivers the intuitive results for c_1 and c_2. Since $\{r_1\}$ is the only proper scenario based on c_1 and $\mathcal{W} \cup Conclusion[\{r_1\}] \vdash B$, the context c_1 entails the formula B, which, you'll recall, says that the object in front of you is blue. And given that $\{r_2\}$ is the only proper scenario based on c_2 and $\mathcal{W}' \cup Conclusion[\{r_2\}] \nvdash B$, the context c_2 doesn't entail B.

With this, our basic defeasible reasoner is complete. We interpret it as a *model reasoner*: If it outputs X in the context c, then it's rational for you to believe (or you should believe) that X in the scenario that c stands for. And if it doesn't output X in c, then it's *not* rational for you to believe (or it's not the case that you should believe) that X in this scenario.[11] Also, I will often represent contexts as *inference graphs*, such as the two graphs depicting c_1 and c_2 in Figure 1. Here's how they should be read: A simple arrow going from a node X to a node Y stands for a vanilla default of the form $\frac{X}{Y}$. A crossed out arrow starting from a node X and pointing to another arrow represent an exclusionary default of the form $\frac{X}{Out(r)}$, with the second arrow

[10] This is one of the two natural ways to define consequence in the present framework. The alternative definition says that X follows from c just in case $\mathcal{W} \cup Conclusion[\mathcal{S}] \vdash X$ for *some* proper scenario $\mathcal{S}$ based on c. Nothing important hinges on which of the two definitions we adopt: Almost all contexts we're going to discuss will have only one proper scenario, and when there's one proper scenario, the definitions lead to the same result. We're going to briefly revisit the alternative definition in footnote 28.

[11] It's natural to think that the model comes with the all-or-nothing picture of doxastic attitudes. However, it turns out to be possible to extend it to capture the degree-of-confidence talk too, at least to some extent—see (Knoks, forthcoming), cf. (Pollock, 1995, 2001, 2010). Why not use the extended model here? Well, first, it's relatively complex. And second, we don't need its expressive power to reach the goals of this paper. In the present context, abstracting away from degrees of confidence is a useful idealization.

standing for the rule r. Notice that some of the nodes are black, while others are gray. The black nodes stand for the (atomic) propositions that follow from the context; the gray ones for the propositions that do not.

3 Capturing conciliationism

Now let's see how the core idea motivating conciliatory views can be captured in our defeasible reasoner. We start by considering a well-known case, where the conciliatory response seems particularly intuitive:

> **Mental Math**. My friend Megan 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 Megan says she got $45. I dramatically reduce my confidence that $43 is the right answer.[12]

Mental Math describes fairly complex reasoning, and we shouldn't miss three of its features: First, we can distinguish two components in it, the mathematical calculations and the reasoning prompted by Megan's announcement. What's more, it seems perfectly legitimate to call the former the agent's *first-order reasoning* and the latter her *second-order reasoning*. Second, the agent's initial confidence in $43 being the correct answer is based on her calculations, and it gets reduced *because* the agent becomes suspicious of them. And third, Megan's announcement provides for a very good reason for the agent to suspect that she may have erred in her calculations.

Bearing this in mind, let's try to capture the agent's reasoning in the model. As a first step, we introduce a new predicate $Seems(\cdot)$ to our language. The formula $Seems(X)$ is meant to say that the agent has reasoned to the best of her ability about whether X, coming to the conclusion that X as a result. It should be clear that the reasoning implied by $Seems(X)$ will depend on X, and, thus, that it will vary from one case to another: If X is a mathematical claim, $Seems(X)$ is the result of doing calculations of the sort described in Mental Math. If X is a philosophical claim, $Seems(X)$ is the result of a careful philosophical investigation. Also, note that $Seems(X)$ is compatible with $\neg X$. Since the agent is fallible, the fact that she has reasoned to the best of her ability doesn't guarantee that her conclusion is correct.

Presumably, though, situations where the agent's best reasoning leads her astray are relatively rare, and so it seems reasonable for her to go by her best reasoning. In the end, she doesn't really have other alternatives. This motivates the following default rule schema:

> **Significance of first-order reasoning**: $r(X) = \dfrac{Seems(X)}{X}$, or if your best first-order (or domain-specific) reasoning suggests that X, conclude X by default.

[12] The case is adopted from (Christensen, 2010) near verbatim. I've given the friend a name.

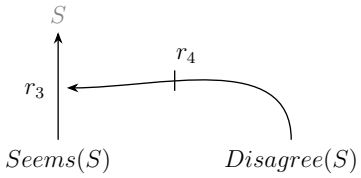

Fig. 2 Mental Math, preliminary

In Mental Math, we'd instantiate this schema with the rule $r_3 = \frac{Seems(S)}{S}$, where S stands for the proposition that my share of the bill is \$43. What Megan's announcement brings into question is exactly the connection between $Seems(S)$ and S. My mental calculations are generally reliable, but I also know that now and then I make a mistake. The announcement, then, suggests that this may well have happened.

To capture the effects of the announcement, we use a second designated predicate $Disagree(\cdot)$. The formula $Disagree(X)$ is meant to express the idea that the agent is in genuine disagreement over X. And I say *genuine* here to distinguish the disagreements that seem to be epistemically significant intuitively from *merely apparent* disagreements, such as verbal disagreements and disagreements based on misunderstandings.[13] So $Disagree(S)$ means that there's genuine disagreement—between Megan and me—over whether or not my share of the bill is \$43. And we capture the effects of this disagreement by means of the default rule $r_4 = \frac{Disagree(S)}{Out(r_3)}$ which says the following: In case there's genuine disagreement about whether S, the rule r_3, that is, the rule that sanctions concluding S on the basis of $Seems(S)$, or the rule that we used to capture the significance of my mental calculations, is to be excluded. Thus, r_4 is what expresses the distinctively conciliatory component of the complex reasoning discussed in Mental Math.

Now notice that we can generalize from r_4 to another default rules schema:[14]

> **Significance of disagreement**: $r'(X) = \frac{Disagree(X)}{Out(r(X))}$, or if there's genuine disagreement over X, stop relying on your first-order reasoning about X by default.

I take this schema to express the core idea motivating conciliatory views—or the core of conciliationism—in our model.[15]

[13] The distinction is standard—see, for instance, (Matheson, 2015b, pp. 7–8).

[14] An anonymous referee points out that it'd be more accurate to have a hierarchy of schemas, as opposed to the first-order reasoning schema and the disagreement schema alone. In a more technical investigation we'd indeed work with a hierarchy of schemas. But nothing hinges on collapsing the hierarchy here.

[15] It's natural to wonder how we might capture the core idea behind steadfast views in the model. I think a steadfast reasoner should be modeled as a reasoner that never relies on the distinctively conciliatory schema $r'(X) = \frac{Disagree(X)}{Out(r(X))}$.

But how shall we model Mental Math? As a first pass, we can try to express it in the context $c_3 = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W} = \{Seems(S), Disagree(S)\}$ and $\mathcal{R} = \{r_3, r_4\}$. (The context is depicted in Figure 2.) It can be verified that S doesn't follow from c_3. But while this is the intuitive result, several important features of Mental Math aren't represented by c_3 adequately. In particular, it misleadingly suggests that the agent doesn't reason to the conclusion that there's genuine disagreement over the amount of the bill, but, rather, starts off knowing that there's such disagreement. Admittedly, the description of the case glances over this component of the reasoning. But it is implied by the background story: ".. Megan and I have been going out to dinner for many years.. we've been right equally often.. neither Megan nor I have had more wine or coffee than usual". This reasoning has its own specific domain, and so it seems natural to capture it using the familiar schema $\frac{Seems(X)}{X}$. The most straightforward thing to do would be to instantiate it with the formula $Disagree(S)$, and, for many contexts, this would be a perfectly fine way to proceed. Here, however, we're going to do something slightly more sophisticated, because it will help us resolve the paradoxes associated with disagreements over peerhood.

Most conciliationists would agree that it's rational for you to back off from your belief in X once both of the following conditions obtain. First, you have an epistemic peer who has carefully thought about X. And second, this peer doesn't believe X.[16] Plausibly, your reasoning toward the conclusion that there's genuine disagreement over X must involve reasoning about each of these conditions. What we do, then, is explicitly distinguish the reasoning about the peer status of the person you're disagreeing with and the reasoning about her take on X. In Mental Math, there's plenty of reason to think that both conditions are satisfied: Megan and I are equals when it comes to mental calculations, and there's very little place to doubt that she doesn't think that my share of the bill is $43.

Bearing this in mind, we let M and MDS stand for the propositions, respectively, that Megan is my peer, and that Megan doesn't think that my share of the bill is $43, and we use the pair of contexts $c_4 = \langle \mathcal{W}, \mathcal{R} \rangle$ and $c_5 = \langle \mathcal{W}', \mathcal{R} \rangle$ to capture the scenario: The first represents the state of affairs before Megan's announcement, the second after. The hard information $\mathcal{W}$ of the first context c_4 comprises the formulas $Seems(S)$, $Seems(M)$, and $(M \& MDS) \supset Disagree(S)$. Notice that the third formula captures the above observation about the conditions that suffice for an epistemically significant disagreement: If Megan is my peer and she does believes that my share of the bill is $43, $M \& MDS$, then she and I are indeed in genuine disagreement over whether my share of the bill is $43, $Disagree(S)$. The hard information $\mathcal{W}'$ of c_5, in turn, extends $\mathcal{W}$ with the formula $Seems(MDS)$, capturing the idea that my best reasoning suggests that Megan doesn't think that my share of the bill is $43. Both contexts share the same set of rules $\mathcal{R}$ which is comprised of the familiar r_3 and r_4, as well as two further instances of the first-order reasoning schema $r_5 = \frac{Seems(M)}{M}$

[16] It may be necessary to hedge these claims to make sure that they apply to the standard cases only. But nothing important seems to hinge on this.

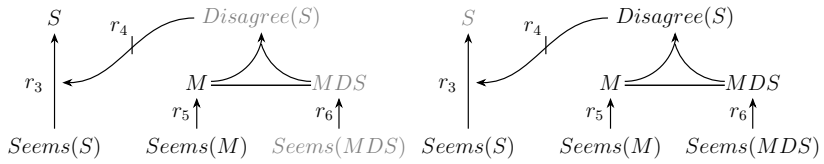

Fig. 3 Mental Math, final, c_4 (left) and c_5 (right)

and $r_6 = \dfrac{Seems(MDS)}{MDS}$. Both contexts are depicted in Figure 3, with the new type of arrow standing for the proposition $(M\&MDS) \supset Disagree(S)$.

It can be verified that $\{r_3, r_5\}$ is the only proper scenario based on c_4, and that, therefore, S and M follow from c_4. So the model suggests that, before Megan's announcement, it's rational for me to believe that Megan is my peer and that my share of the bill is \$43. Moving on to c_5, there's again a unique proper scenario based on it, namely, $\{r_4, r_5, r_6\}$, which, in turn, implies that M, MDS, and $Disagree(S)$ follow from c_5, while S doesn't. So the model suggests that, after Megan's announcement, it's rational for me to believe that Megan is my peer; that she doesn't think that my share of the bill is \$43; that she and I are in genuine disagreement over whether my share of the bill is \$43, as well as that it's no longer rational for me to believe that my share of the bill is \$43. Thus, our model delivers the intuitive result.

Before we turn to scenarios involving higher-order disagreements, it'll be instructive to take a look at a variation on Mental Math which is often brought up against conciliatory views:[17]

> **Careful Checking**. I consider my friend Megan 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. Megan announces that she got \$45.[18]

Unlike it was in Mental Math, here it doesn't seem right for me to lose confidence in my answer—it's extremely unlike that someone in my situation could have gotten the same wrong answer each time. And yet, says the objector, the conciliatory views do suggest that I don't believe that my share of the bill is \$43.

[17] See, for instance, (Lackey, 2010a,b).

[18] The case is adopted from (Christensen, 2011, p. 8), again, near verbatim.

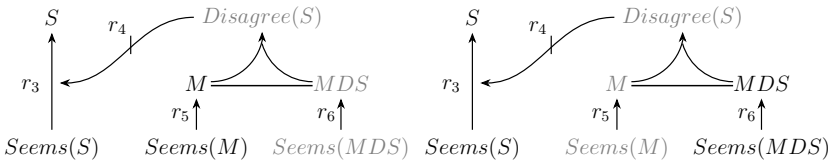

Fig. 4 Careful Checking, two versions, c_4 (left) and c_6 (right)

How would we represent Careful Checking in our model? Well, I contend that the scenario is underdescribed, and that, therefore, there isn't just one, but two ways to capture it formally: Given some ways of filling in the details, we'll want to opt for one context; given others, we'll want to opt for another one. Before we turn to these contexts, however, let me emphasize that, in Careful Checking, it seems very reasonable for the agent to not take her disagreement with Megan to be epistemically significant. David Christensen (2011) suggests this much as well, adds that there's actually good reason for the agent to suspect that something screwy is going on with Megan, and lists a number of possible explanations: She might be playing a joke on the agent, deliberately making false claims, be drunk, tripping, exhausted, or experiencing a bizarre mental malfunction.[19]

Now let's suppose that the most probable explanation of Megan's announcement is that she's trying to play a joke on me—and it should be clear that on some ways of filling in the details of the scenario this explanation will be the most probable. If so, then it's no longer the case that my best (first-order) reasoning suggests that Megan doesn't think that my share of the bill is \$43. And given that this is just what the formula $Seems(MDS)$ says, we can conclude that the context representing this version of Careful Checking shouldn't have it in its hard information. Notice that subtracting $Seems(MDS)$ from c_5 results in the familiar context c_4, that is, the context that we used to represent the first episode of Mental Math, and it makes for a fine representation of the particular version of Careful Checking too. To take another example, suppose that the scenario's details are filled in in such a way that other explanations from Christensen's list becomes more probable, say, that Megan is tripping or experiencing a bizarre mental malfunction. If so, my best reasoning *will* suggest that Megan doesn't think that my share of the bill is \$43, and so $Seems(MDS)$ will have to be part of the representation. However, here my best (first-order) reasoning won't any longer suggest that Megan is my epistemic peer—she loses this status temporarily—and so the formula $Seems(M)$ must be left out of the representation. We can capture this version of Careful Checking in the context c_6 which is just like c_5, except for the fact that the formula $Seems(M)$ is absent from the hard information—see Figure 4 where c_6 is depicted alongside c_4. My more general claim, then, is this: On every way of spelling out the details of Careful Checking, either c_4 or c_6 will serve as an adequate representation. We already know that the only proper scenario based on c_4 is $\{r_3, r_5\}$, and that S does while $Disagree(S)$ doesn't follow from it. And it can be verified—which we won't do here—that the only proper scenario based on c_6 is

[19] See (Christensen, 2011, p. 9).

$\{r_3, r_6\}$, and that, therefore, S does while $Disagree(S)$ doesn't follow from c_6 too. So on either representation, we get the intuitive result that it's rational for me to retain my original belief about the amount of the bill and *not* to believe that there's genuine disagreement over it. The case doesn't undermine conciliationism.[20]

Summarizing, our model suggests the following take on conciliationism: It's not the simple theory saying that you're invariably required to give up your belief in X as soon as you're in disagreement over X with an epistemic peer—or, perhaps, as soon as it's rational for you to think that you're in such disagreement. Instead, it's a theory that's more complex and more structured, and what it says runs roughly as follows: If your best first-order (or domain-specific) reasoning suggests that X and it's rational for you to believe that you're in disagreement over X with an epistemic peer, then, under normal circumstances, you should bracket your first-order reasoning about X and avoid forming beliefs on its basis.[21]

4 Two paradoxes of epistemic peerhood

Mulligan (2015) presents three paradoxes that arise when conciliationism attempts to deal with disagreements over epistemic peerhood. The first two are illustrated using the same case:

> **Disagreement over Peerhood**. Imagine that I disagree with my friend Francis about the truth of some proposition X. I believe that X is true and Francis believes that X is false. Since I regard Francis as my epistemic peer with

[20] One might be suspicious of the way we handled Careful Checking, wanting to ask the following question: 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? Well, first off, this move allow us to handle Careful Checking, but it is of little help with many other cases, including many cases involving higher-order disagreements. And second, the feeling of suspicion may have more to do with the idealizations of the model, and less with the general idea for how to handle the scenario: Intuitively, my calculations make for a very good reason for me to believe that my share of the bill is $43 and Megan's announcement too makes for a good reason to believe that there's genuine disagreement between us, although a weaker one. The context c_4, by contrast, may suggest that there's absolutely no reason for me to believe that there's genuine disagreement in the case at hand. This, however, is only because the model has no place for degrees of confidence. A more expressive model would let us talk about the reasoning agent's (relative) degrees of confidence and express such claims as "I'm more confident that my share of the bill is $43 than that Megan really doesn't think that it is". The additional expressive power would let us represent the case more fully and distinguish it from the first episode of Mental Math. For more on this, see (Knoks, forthcoming).

[21] Having seen how the core of conciliationism can get captured in default logic, one might still have a lingering worry: While conciliatory views are concerned with the question of how one's doxastic attitudes should *change over time* (in response to disagreements), default logic doesn't really capture any timeline—certainly, not in a natural way. So how can default logic be a well-suited framework for modeling conciliationism? In response, even though there's a tendency to construe conciliatory views in diachronic terms—especially, among those who are attracted to probabilistic models of belief update—some of the best-known advocates of conciliatory views think of them in purely synchronic terms, or, roughly, as views that aren't about how one is to update one's doxastic attitudes, but about which doxastic attitudes are rational to have, given the evidence at one's disposal. On this way of thinking, questions about timeline and the order in which evidence is processed are not important (or relatively unimportant), and so default logic presents itself as a very natural framework to model conciliationism. Thanks to an anonymous referee for pressing me to address this worry.

> respect to X, I revise my confidence in X 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... [am convinced] that Richard is my epistemic peer with respect to assessments of epistemic peerhood.[22]

The first paradox has to do with what doxastic attitude conciliationism recommends to take toward X. It seems intuitive to reason as follows. Being a committed conciliationist, I should dramatically reduce my confidence in the proposition that Francis is my epistemic peer in response to Richard's announcement. As a result, my disagreement with Francis should lose its epistemic significance and my initial confidence in X should be restored. However, Mulligan sees a deep problem here: There are many instances of X for which restoring the initial level of confidence will seem counterintuitive. For, in addition to making me reduce confidence in Francis' peerhood, Richard's announcement also makes me realize that my ability to assess peerhood is more flawed than I took it to be. This, in turn, can cast doubt on my initial judgment of X. In fact, it *should* cast doubt on my initial judgment in case this judgment relied on the ability. Thus, Mulligan writes, "..To elicit the paradoxical result, X must be chosen such that mistakes in assessing peerhood clearly cast doubt on my ability to get X right. When there is a clear connection between X and my ability to assess my epistemic peer relations, the rational response is to call into question my own epistemic faculties and downgrade my confidence in X. A good example for X here is *Vanessa will get an "A" on her philosophy exam*. If I discover that I am worse at assessing exposure to the evidence, intelligence, and freedom from bias than I thought I was, then I am worse at assessing Vanessa's ability to get an "A" " (Mulligan, 2015, p. 70). The first paradox, then, is this: Instantiate X in Disagreement over Peerhood with *Vanessa will get an "A" on her philosophy exam*. On the one hand, conciliationism seems to recommend that I retain full confidence in this proposition. On the other, there's the intuition that my confidence should be much lower.

As a first step toward a solution, let's express the scenario in our model. Let V stand for the proposition that Vanessa will get an "A" on her philosophy exam; F and FDV for the propositions, respectively, that Francis is my epistemic peer and that Francis doesn't think that Vanessa will get an "A"; and R and RDF for the propositions, respectively, that Richard is my epistemic peer and that Richard doesn't think that Francis is my peer.[23] We use the context $c_7 = \langle \mathcal{W}, \mathcal{R} \rangle$ to represent the scenario. Its hard information $\mathcal{W}$ consists of the formulas $Seems(V)$, $Seems(F)$, $Seems(FDV)$, $Seems(R)$, $Seems(RDF)$, as well as two material conditionals specifying the conditions that are jointly sufficient for epistemically significant disagreements, $(F\&FDV) \supset Disagree(V)$ and $(R\&RDF) \supset Disagree(F)$. The context's set of rules $\mathcal{R}$ contains five instances of the first-order reasoning schema, $r_7 = \dfrac{Seems(V)}{V}$, $r_8 = \dfrac{Seems(F)}{F}$, $r_9 = \dfrac{Seems(FDV)}{FDV}$, $r_{10} = \dfrac{Seems(R)}{R}$,

[22] The case is adopted from (Mulligan, 2015, p. 69) near verbatim. I've changed the propositional letter from P to X and given the case a name.

[23] I'm cutting corners here: The formulas F and R should actually be relativized to subject matter. Thus, F, for instance, should say that Francis is my epistemic peer with respect to assessing V.

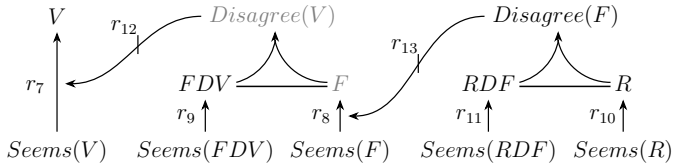

Fig. 5 Disagreement over Peerhood

and $r_{11} = \dfrac{Seems(RDF)}{RDF}$, as well as two instances of the disagreement schema, namely, $r_{12} = \dfrac{Disagree(V)}{Out(r_7)}$ and $r_{13} = \dfrac{Disagree(F)}{Out(r_8)}$. The context is depicted in Figure 5.

The unique proper scenario based on c_7 is $\{r_7, r_9, r_{10}, r_{11}, r_{13}\}$. The rules r_{10} and r_{11} let the defeasible reasoner conclude that there's genuine disagreement over Francis' peer status, $Disagree(F)$, which, in turn, precludes it from using the rule $r_8 = \dfrac{Seems(F)}{F}$. As a result, the reasoner never reaches the conclusion that there's genuine disagreement over Vanessa's grade, $Disagree(V)$, leaving the rule $r_9 = \dfrac{Seems(FDV)}{FDV}$ intact. This means that V and $Disagree(F)$ do, while F and $Disagree(V)$ do not follow from c_7. Thus, the model suggests that I should believe that Vanessa is going to get an "A" and that there's genuine disagreement over Francis' peerhood, and that I should *not* believe that Francis is my peer. This is roughly what the recommendations of conciliationism are, according to Mulligan.

Now notice that our formalization compartmentalizes the reasoning in the scenario, letting us see what exactly the commitments of conciliationism are. Mulligan characterizes the case as one where I learn, as a result of Richard's announcement, that I committed two connected errors: First, I judged Francis to be my epistemic peer, which he turned out not to be, and, second, I downgraded my initial confidence in my prediction of Vanessa's performance—see (Mulligan, 2015, p. 70ff). The model suggests that this is a mischaracterization. Contra Mulligan, Richard's announcement doesn't conclusively show that I made an error in assessing Francis' peerhood, but only that there's good reason to suppose that I may have. (Here things aren't any different from any old case of disagreement: As long as the only thing I know is that I'm in disagreement over X with an epistemic peer, there's no way of telling which one of us has erred.) Relatedly, the model suggests that my disagreement with Richard stops short of making me lose confidence in my ability to assess peerhood. Instead, its epistemic effects are confined to making me bracket the reasoning that led me to conclude that Francis is my peer, which, in turn, makes my disagreement with Francis lose its epistemic significance. In light of this, the recommendation to revert to the original confidence in Vanessa's ability to get an "A" seems perfectly intuitive.

It's worth being explicit about where exactly my analysis of Disagreement over Peerhood parts ways with Mulligan's: He thinks that conciliatory reasoning issues in two (inconsistent) recommendations: First, there's the recommendation that I re-

vert to my initial confidence in Vanessa's ability to get an "A". Second, there's the recommendation that my confidence in Vanessa's ability to get an "A" be lower than that. On my analysis, by contrast, conciliatory reasoning issues only in the first recommendation. Mulligan thinks that the second recommendation stems from my loss of confidence in my capacity to assess people's intelligence. On my analysis, I never lose confidence in my capacity to assess people's intelligence. Mulligan thinks that I lose confidence in this capacity because, upon hearing Richard's announcement, I discover that I erred in my assessment of Francis. On my analysis, upon hearing Richard's announcement, I do not discover that I erred in my assessment of Francis, but only gain a good reason to suspect that I may have erred. Mulligan thinks that discovering that you're in disagreement over X with an epistemic peer can make you learn that you erred about X. I think that it cannot, and that conciliationism doesn't entail that it can.

But one could grant all of this and still worry that a scenario in the vicinity of Disagreement over Peerhood can revitalize Mulligan's paradox, namely, a scenario in which I do in fact learn that I erred in judging Francis to be my peer. I have two things to say in response here. First, I suspect that it's going to be very difficult to come up with a fleshed-out description of this case, generating the intuitions needed for the paradox. Just think of what has to happen for me to *learn* that Francis turns out *not* to be my peer when it comes to predicting people's ability to do well on philosophy exams. Presumably, the one thing that will matter is how successful our predictions are. And it's very tempting to think that I shouldn't take a single successful prediction on my part to be decisive. (Couldn't it be a matter of pure luck that, in this one case, I got it right and Francis didn't?) So I'd probably need to compare our track records and see that my predictions turn out to be correct more often than his. What's more, it's worth thinking about what assessing predictions of this sort involves: Suppose Francis' prediction regarding Vanessa turned out to be wrong. One consequence of this is that Vanessa actually does end up getting an "A", which is just what I thought was going to happen and which strongly suggests that my judgment of her abilities was correct. All in all, then, I've trouble seeing how learning that I've erred about Francis' peerhood could make me lose confidence in my ability to judge exposure to evidence, intelligence, and freedom from bias.

My second point is this: Even if it's possible to come up with a case of the sort Mulligan needs, the formal model points to a natural way to accommodate the conflicting intuitions. For concreteness, let's consider the following scenario, bracketing the worries raised in the previous paragraph and pretending that my doxastic responses in it strike us as rational:

> **Twinessa**. The story goes like in Disagreement over Peerhood, with *Vanessa will get an "A"* taking the place of X. Now, however, Vanessa has a twin, Twinessa, who is taking the same philosophy exam. I judge Twinessa to be as gifted as her sister, predicting that she will get an "A". Francis too judges the two sisters to be equally gifted, predicting that Twinessa will not get an "A". Then the exam results come in, and I learn that Twinessa has gotten an "A". This makes me realize that Richard was right, and that I erred when I judged Francis to be my peer. As a result, I lose confidence in my ability to

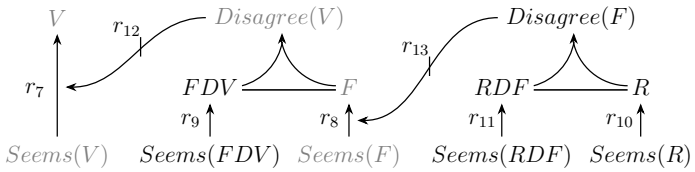

Fig. 6 Twinessa

> assess exposure to evidence, intelligence, and freedom from bias, as well as dramatically reduce my confidence in *Vanessa will get an "A"*.

Notice that our challenge is to reconcile the recommendation of conciliationism to retain full confidence in *Vanessa will get an "A"* and my actual response. And I think that there's a perfectly natural way to do it: We must distinguish between two episodes in the scenario, each captured by a different context. The first one pauses the story just before Twinessa's exam results are revealed. Leaving the part of the story that concerns Twinessa unrepresented (purely for the sake of convenience), we can capture the scenario using the same context c_7 that we used to capture Disagreement over Peerhood. The formula V follows from it, supporting the recommendation of conciliationism. The second episode is the endpoint of the story. Leaving Twinessa unrepresented, we can capture it using the context $c_8 = \langle \mathcal{W}, \mathcal{R} \rangle$ which is just like c_7, except for its hard information $\mathcal{W}$ doesn't contain the formulas $Seems(V)$ and $Seems(F)$—see Figure 6 for a picture.[24] Why is excluding these formulas justified? Well, recall that $Seems(V)$ says that I have reasoned to the best of my ability about whether Vanessa will get an "A" and come to the conclusion that she will. But, clearly, this is no longer true at the end of the scenario: I've lost all confidence in my ability to assess exposure to evidence, intelligence, and freedom from bias, and so I won't be making any predictions. As for $Seems(F)$, it's a stipulation of the scenario that I learn that Francis is not my peer.

It can be verified that V doesn't follow from c_8 which fits with the story. And, with my best reasoning about Vanessa no longer supporting the conclusion that she will get an "A" and my reasoning about Francis no longer supporting the conclusion that he's my peer, conciliationism no longer issues any recommendations. There's no tension, no paradox.

The second paradoxical consequence of conciliationism that Disagreement over Peerhood is said to elicit is similar to the first: Once I realize that I've made an error in assessing Francis' peerhood, I learn that I'm worse at assessing peerhood than I thought. This makes me lose confidence in my ability to assess peerhood, and, thereby, question my initial judgment of Richard. Here Mulligan writes, ".. Why should I have unchanged confidence that Richard is.. my epistemic peer? I ought

[24] The representation assumes that my loss of confidence in my ability to assess peerhood leaves my first-order reasoning about Richard intact. This might be disputed—and it's roughly what Mulligan's second paradox revolves around—but nothing hinges on this: We could easily leave out the formula $Seems(R)$ from the hard information too.

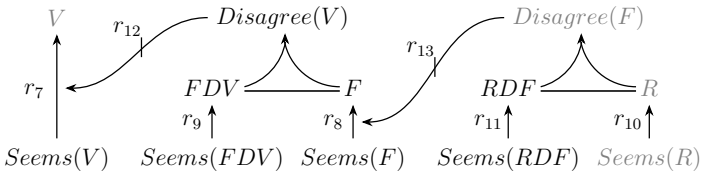

Fig. 7 Francine

to be less confident that Richard is my epistemic peer and thus more confident that Francis is my epistemic peer (since Francis's peerhood was undermined by Richard's opinion). When I learn of Richard's disagreement about the proposition *Francis is my epistemic peer with respect to* X, conciliationism simultaneously demands that (1) I lose confidence in *Francis is my epistemic peer with respect to* X, and (2) I lose confidence that I am a good judge of peerhood, thus losing confidence that Richard is my epistemic peer and gaining confidence in the proposition *Francis is my epistemic peer with respect to* Q, which contradicts (1)" (Mulligan, 2015, p. 73).

Given the preceding discussion, it should be clear how this tension can be resolved. First off, we should say that conciliationism only demands that I bracket the reasoning that depends on *Francis is my epistemic peer* in response to my disagreement with Richard, and that it does *not* demand that I lose confidence in my ability to assess peerhood. The fact that Richard and I disagree doesn't by itself show that my initial judgment of Francis was mistaken. There's no paradox here.

And while one might again worry that a scenario where I do learn that my judgment of Francis was mistaken—some scenario like Twinessa—could revitalize the paradox, both of the above considerations again apply. First, it won't be easy to flesh out the details of the story, while maintaining the intuition that I should lose confidence in my abilities. (Drawing conclusion on the basis of one mistaken prediction seems reckless. What's more, Francis' turning out not to be my peer implies that Richard was right, and, hence, that I assessed Richard's peerhood adequately. Plausibly, this fact should offset the loss of confidence in my abilities prompted by the mistaken judgment about Francis.) And second, even if there was such a story, the model suggests a natural way to accommodate the conflicting intuitions about it. We could describe a case—structurally similar to Twinessa—where Francis has a twin, Francine, and where, upon learning that Francine isn't my peer, I lose confidence in my ability to assess peerhood, with the effect that I no longer trust my assessment of Richard, but, somewhat surprisingly, still trust my assessment of Francis.[25] We can, again, distinguish between two episodes in this story. The first pauses it just before I learn about having erred regarding Francine. At this point, conciliationism recommends that I bracket the reasoning based on *Francis is my epistemic peer*. If we leave

[25] This is less absurd than it sounds at first. Plausibly, I can be more (rationally) confident of my assessment of the peerhood of one person than another. So it could well happen that I'm more confident in the (first-order) reasoning suggesting that Francis is my peer than in the reasoning suggesting that Richard is my peer. And it doesn't seem outlandish to hold that finding out that I erred about Francine will affect the less entrenched reasoning that led me to conclude that Richard is my peer, but not the more entrenched reasoning that led me to conclude that Francis is my peer.

the reasoning concerning Francine out of the picture, we can capture this episode using the familiar context c_7—see Figure 5. The second episode is, again, the endpoint of the story. Here I no longer judge Richard to be my epistemic peer, which is enough for my disagreement with him to cease to be epistemically significant for me. And this, in turn, means that there's nothing precluding me from relying on my (first-order) reasoning about Francis' peerhood, with the result that it has its regular downstream effects—that is, that I believe that Francis is my peer, and that I don't believe that Vanessa will get an "A". This episode can be captured by the context $c_9 = \langle \mathcal{W}, \mathcal{R} \rangle$ which is like c_7, with the exception that its hard information doesn't contain the formula $Seems(R)$—see Figure 7 for a graph. Thus, there's no contradiction between Mulligan's (1) and (2). These demands apply at different time points in the story.

So both paradoxes can be resolved with the help of our model. Conciliationism comes out unscathed.

5 A third paradox and self-undermining chains of thought

The third paradox has to do with the following scenario:

> **Modest Jimmy**. Let's imagine that I am, following conciliationism, factoring in the opinion of Jimmy, whom I believe to be my epistemic peer on some matter X about which we disagree. What is the rational response upon discovering that Jimmy disagrees about our epistemic parity? Imagine that Jimmy says this: ".. I know that I disagree with you about all kinds of things, including X, but I'm not your intellectual equal. You have good reason to ignore my opinions and stick to your guns."[26]

Mulligan argues that this scenario reveals another impossible tension in conciliationism, reasoning about it roughly as follows. On the one hand, conciliationism demands that I lower my confidence in the proposition *Jimmy is my epistemic peer*. On the other, once my confidence in *Jimmy is my epistemic peer* has gone down, I have reason to discount Jimmy's opinions, including his opinion about our peerhood. And so my confidence in my original judgment, *Jimmy is my epistemic peer*, should go up.

Unfortunately, the considerations that helped us resolve the previous paradoxes have no force here: While we could point out that Jimmy's announcement doesn't conclusively show that I erred in assessing his peerhood, it wouldn't bring us any closer to resolving the paradox. This time, the puzzle has nothing to do with the changes in how confident I am in my ability to assess peerhood, but, rather, with the epistemic significance of my disagreement with Jimmy itself.

Now let's see what the formal model has to say about Modest Jimmy—for simplicity, we leave the disagreement over X unrepresented. Let J stand for the proposition that Jimmy is my epistemic peer and JDJ for the proposition that Jimmy doesn't think that he's my peer. The scenario can then be expressed as the context $c_{10} = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ is comprised of the formulas $Seems(J)$, $Seems(JDJ)$, and

[26] (Mulligan, 2015, p. 75). As before, X has been substituted for P.

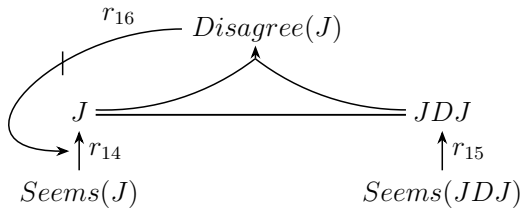

Fig. 8 Modest Jimmy

$(J\&JDJ) \supset Disagree(J)$, while $\mathcal{R}$ is comprised of the rules $r_{14} = \dfrac{Seems(J)}{J}$, $r_{15} = \dfrac{Seems(JDJ)}{JDJ}$, and $r_{16} = \dfrac{Disagree(J)}{Out(r_{14})}$, with the first two instantiating the first-order reasoning schema and the third the disagreement schema. The context is depicted in Figure 8.

It turns out that there are no proper scenarios based on c_{10}.[27] And this looks like pretty bad news for conciliatonism: On our definition of consequence, a formula X follows 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 based on a context, any formula satisfies the right-hand side of the biconditional vacuously and thus follows from the context. So c_{10} entails all formulas, suggesting that Modest Jimmy may indeed reveal some sort of incoherence in conciliatory views.[28]

But the issue this scenario makes manifest is actually not specific to conciliationism, and it is not as problematic as may seem. To see this, let's consider the following case, adopted from John Pollock:

> **Pink-Elephant Disorder**. People generally tell the truth. However, some people suffer from a malady know as "pink-elephant disorder". In the presence of pink elephants, they become strangely disoriented so that their statements about their surroundings cease to be reliable. Now imagine Robert, who tells us that the elephant beside him looks pink. In ordinary circumstances, we would infer that the elephant beside Robert does look pink, and hence probably is pink. However, Robert suffers from pink-elephant disorder. So if it were

[27] This can be verified by enumeration, going through every subset of $\mathcal{R}$. Or we can save ourselves some time by realizing that any proper scenario based on c_{10} would have to include r_{15}. For this rule is triggered by the hard information and nothing interferes with it. This leaves us with four candidate scenarios, namely, $\{r_{15}\}$, $\{r_{14}, r_{15}\}$, $\{r_{15}, r_{16}\}$, and $\{r_{14}, r_{15}, r_{16}\}$. The first fails to qualify as proper because it doesn't include r_{14}, which is triggered and not excluded in it; the second because it doesn't include r_{16}, which is triggered and not excluded in it; the third because it includes r_{16}, which isn't triggered in it; and the fourth because it includes r_{14}, which is excluded in it.

[28] When defining consequence, I stated an alternative definition in a footnote: X follows from a context c if and only if X follows from *some* proper scenario based on c. It's natural to wonder if it might be possible to avoid the problem by switching to this definition. On it, nothing follows from c_{10}, which may look like an improvement—in the end, it's not implausible that the right thing to do in the scenario is to suspend judgment on Jimmy's peerhood. However, nothing would be gained by the switch: The alternative definition comes with a recommendation for *universal suspension*. Even the formula $Seems(JDJ)$ doesn't follow from the context.

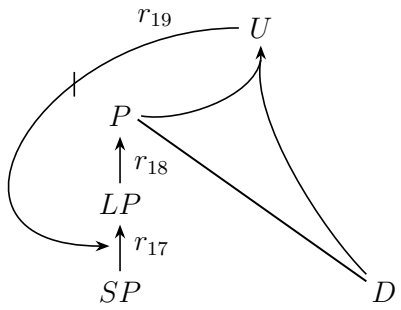

Fig. 9 Pink-Elephant Disorder

> true that the elephant beside Robert is pink, we could not rely upon his report to conclude that it is. So we should not conclude that it is pink.[29]

Pink elephants notwithstanding, this case illustrates a perfectly familiar and sensible sort of reasoning: We sometimes start off from seemingly compelling considerations and then, after a longer chain of thought, realize that these considerations weren't actually safe to rely on. Our usual reaction in such cases is to simply stop relying on these considerations. However, were we to reason about these cases in the way Mulligan reasons about Modest Jimmy, we'd be concluding, instead, that there's an impossible tension in our reasoning. To see this, simply apply his line of thought to Pink-Elephant Disorder: On the one hand, I must lower my confidence in the proposition *Robert is reliable* after Robert's announcement. On the other, once my confidence in *Robert is reliable* has gone down, I have reason to discount Robert's opinion, including his opinion about there being a pink elephant beside him. But if my confidence in *There's a pink elephant beside Robert* is low, my confidence in my original judgment, *Robert is reliable*, should go up.

Thus, we get a paradox where there should be none, and this suggests that the original problem doesn't lie with conciliationism, but, rather, with the way we—following Mulligan—have been thinking about the case. We have been relying on a very simple conception of the dependence relations between the various considerations at play in Modest Jimmy, and, in particular, we have been assuming that my disagreement with Jimmy (regarding our epistemic peerhood) loses all of its epistemic significance for me once my confidence in *Jimmy is my epistemic peer* goes down. I think that this assumption is mistaken, and that it is what gave rise to the problem. I also think that what's actually going on in Modest Jimmy is this: My disagreement with Jimmy retains its significance—in spite of the low confidence in *Jimmy is my epistemic peer*—precluding my confidence in *Jimmy is my epistemic peer* from going up, just like my conclusion that Robert is unreliable retains its significance—in spite of the low confidence in *There's a pink elephant beside Robert*—precluding me from relying on Robert's testimony.

[29] See (Pollock, 1995, p. 120) and (Pollock, 2009, pp. 181–2).

A question remains, however: If Modest Jimmy doesn't cause trouble for conciliationism, why does our formal model lead to the problematic result, letting us derive all formulas from c_{10}? The short answer here is that the result obtains because of a bug in the model. A longer answer involves saying that this bug is not specific to our model, but is a general problem for any version of default logic, and that it has to do with the structural relation between the rules in c_{10}. To see that the problem is not specific to Modest Jimmy, we only need to run our defeasible reasoner on Pink-Elephant Disorder—which, we noted, describes a sensible and familiar sort of reasoning.[30] Let SP, LP, and P stand for the propositions, respectively, that Robert says that the elephant beside him looks pink, that the elephant beside Robert looks pink, and that the elephant beside Robert is pink; and let D and U stand for the propositions, respectively, that Robert is suffering from the pink-elephant disorder and that Robert is unreliable. The scenario can, then, be encoded in the context $c_{11} = \langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ consists of the formulas SP and $(P\&D) \supset U$, expressing the fact that Robert becomes unreliable in the presence of pink elephants due to his disorder, and $\mathcal{R}$ contains the rules $r_{17} = \dfrac{SP}{LP}$, $r_{18} = \dfrac{LP}{P}$, and $r_{19} = \dfrac{U}{Out(r_{17})}$. The first two rules instantiate sensible, albeit defeasible, reasoning principles, namely, that people generally tell the truth, and that things that look pink normally are pink; the third expresses the idea that we shouldn't trust Robert in case he is unreliable. The context c_{11} is depicted in Figure 9. Now, just like it was with c_{10}, there are no proper scenarios based on c_{11}, meaning that c_{11} too entails all formulas.

What's at the root of the problem then? Well, if we look at the graphs representing c_{10} and c_{11}, it's fairly easy to identify their problematic components. In the case of c_{10}, it's the rules r_{14} and r_{16}: The former puts the latter in place and, thereby, undermines its own support. In the case of c_{11}, it's all three rules: The rule r_{17} puts r_{19} in place (via r_{18}), and, thereby, too undermines its own support. Each set of rules forms a *vicious cycle*, and it's just a general and well-known fact about default logic that it's poorly suited to handle contexts containing such cycles.[31]

Unfortunately, there's no straightforward fix to our default-logic-based reasoner that would let it handle c_{10}, c_{11}, and other contexts containing cycles. However, it's possible to restate it in the more general framework of *formal argumentation theory*, and, once that's done, a simple tweak to it results in a more sophisticated reasoner that handles contexts containing cycles more adequately.[32] Described at a high level, it disregards all default rules that form vicious cycles and draws conclusions on the basis of only those rules that neither partake in such cycles, nor in any way depend on them. The analogue of the proper scenario based on c_{11} is the empty set, implying that the only informative formulas that follow from c_{11} are SP and D, standing for

[30] In fact, Pollock (2009) presents this scenario as an intuitive counterexample to Reiter's (1980) default logic, an ancestor of the logic we set up in Section 2.

[31] See, e.g., (Horty, 2012, pp. 59–61) for a discussion.

[32] See (Knoks, 2020, forthcoming) for details. Since the seminal work of Dung (1995), it's well-known that various formalisms modeling defeasible reasoning can be represented in formal argumentation theory. Here's what this means, roughly: The logic we defined in Section 2 determines the consequence set of any given context. A representation of this logic in formal argumentation will operate on contexts in an entirely different fashion, but it will map them into the same consequence sets. Also, the simple tweak I mention in the text consists in substituting *preference semantics* for *stability semantics*.

the propositions that Robert says that there's a pink elephant beside him and that he has pink-elephant disorder—which seems perfectly intuitive. And the analogue of the proper scenario based on c_{10} is the singleton $\{r_{16}\}$, implying that JDJ does, while J and $Disagree(J)$ do not follow from the context. So a technically (but not conceptually) enhanced version of our model suggests the following. In Modest Jimmy, it's rational for me to believe that Jimmy doesn't think that he is my peer, and it is *not* rational for me to believe that Jimmy is my peer, nor that there's genuine disagreement over whether he is. This, again, seems perfectly coherent and intuitive.[33]

Summarizing, Mulligan's way of thinking about Modest Jimmy overgeneralizes, leading to trouble in scenarios that have nothing to do with conciliatory views. This suggests that the paradoxical result isn't due to some flaw in conciliationism, but a completely different phenomenon and mistaken assumptions in thinking about cases where this phenomenon manifests itself. What's more, there's a model within a hand's reach that can adequately handle Modest Jimmy. This seems to me to be enough to take Mulligan's third paradox to be resolved.[34]

6 Conclusion

This paper had two goals. The first was to show how the core idea behind conciliatory views can be expressed in a defeasible logic framework, resulting in an intuitive formal model that's useful for understanding how these views behave in cases involving higher-order disagreements. The second goal was to resolve three apparent paradoxes stemming from disagreements over epistemic peerhood. Both of these goals have now been met. With the help of the model, we saw that conciliationism doesn't actually require that one lose confidence in one's ability to assess peerhood in the scenario used to elicit the first two paradoxes, and that this suffices to block them. What's more, we saw that there's good reason to doubt that any scenario in the vicinity might revitalize the paradox: First, once we start filling in the details, it's difficult to maintain the intuition that one should, indeed, lose confidence in one's ability. Second, even if the details could be filled out while maintaining this intuition, the model points to a natural way to accommodate the conflicting intuitions driving the two paradoxes. As for the third paradox, we saw that the phenomenon which gave rise to it isn't specific to conciliationism, and that the seemingly puzzling cases where it manifests itself aren't all that problematic if one keeps track of the dependence relations between the various considerations at play.

[33] When applied to the context capturing the entire story recounted in Modest Jimmy, the model suggests that I should believe that Jimmy doesn't think that the first-order proposition X is true and that I shouldn't believe X myself. So the epistemic significance of my disagreement with Jimmy over X retains its epistemic significance. My intuitions about what attitudes I should take toward X aren't particularly clear, but this recommendations certainly doesn't strike me as counterintuitive, let alone paradoxical.

[34] It's worth adding that the argumentation-theory-based model doesn't only let us deal with Modest Jimmy and Pink-Elephant Disorder adequately, but can also be used to respond to the worries about the behavior of conciliatory views in cases involving the other type of higher-order disagreements, that is, disagreements over conciliatory views themselves—see my (Knoks, forthcoming) for details.

Acknowledgements I'd like to thank the audiences at various venues in Amsterdam, Berlin, College Park, New York, and Palo Alto for feedback on earlier versions of the material presented here. I owe special thanks to John Horty, Eric Pacuit, James Pryor, as well as the two anonymous referees of this journal.

References

Bogardus, T. (2009). A vindication of the equal-weight view. *Episteme*, 6(3), 324–35.

Christensen, D. (2007). Epistemology of disagreement: The good news. *Philosophical Review*, 116, 187–217.

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

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.

Christensen, D. (2021). Akratic (epistemic) modesty. *Philosophical Studies*, 178, 2191–214.

Decker, J. (2014). Conciliation and self-incrimination. *Erkenntnis*, 79, 1099–134.

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.

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.

Fleisher, W. (2021). How to endorse conciliationism. *Synthese*, 198, 9913–39.

Gelfert, A. (2011). Who is an epistemic peer? *Logos and Episteme*, 2(4), 507–14.

Horty, J. (2012). *Reasons as Defaults*. Oxford University Press.

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.

King, N. (2012). Disagreement: What's the problem? Or a good peer is hard to find. *Philosophy and Phenomenological Research*, 85(2), 249–72.

Knoks, A. (2020). *Defeasibility in Epistemology*. PhD thesis, University of Maryland, College Park. doi.org/10.13016/wueu-csc6.

Knoks, A. (Forthcoming). Conciliatory reasoning, self-defeat, and abstract argumentation. *Review of Symbolic Logic*.

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.

Littlejohn, C. (2013). Disagreement and defeat. In D. Machuca (Ed.), *Disagreement and Skepticism* (pp. 169–92). Routledge.

Littlejohn, C. (2020). Should we be dogmatically conciliatory? *Philosophical Studies*, 177, 1381–98.

Matheson, J. (2015a). Are conciliatory views of disagreement self-defeating? *Social Epistemology*, 29(2), 145–59.

Matheson, J. (2015b). *The Epistemic Significance of Disagreement*. Pelgrave Macmillan.

Matheson, J. (2018). Disagreement and epistemic peers. In *Oxford Handbooks Online*. DOI: 10.1093/oxfordhb/97801999353.

Mulligan, T. (2015). Disagreement, peerhood, and three paradoxes of conciliationism. *Synthese*, 192, 67–78.

Pittard, J. (2015). Resolute conciliationism. *The Philosophical Quarterly*, 65(260), 442–63.

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. (2009). A recursive semantics for defeasible reasoning. In G. Simari & I. Rahwan (Eds.), *Argumentation in Artificial Intelligence* (pp. 173–97). Springer.

Pollock, J. (2010). Defeasible reasoning and degrees of justification. *Argument and Computation*, 1(1), 7–22.

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

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

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.