Philos Stud (2017) 174:163–183
DOI 10.1007/s11098-016-0675-6

# Deference, respect and intensionality

**Anna Mahtani**[1]

Published online: 23 April 2016


**Abstract** This paper is about the standard Reflection Principle (van Fraassen in J Philos 81(5):235–256, 1984) and the Group Reflection Principle (Elga in Nous 41(3):478–502, 2007; Bovens and Rabinowicz in Episteme 8(3):281–300, 2011; Titelbaum in Quitting certainties: a Bayesian framework modeling degrees of belief, OUP, Oxford, 2012; Hedden in Mind 124(494):449–491, 2015). I argue that these principles are incomplete as they stand. The key point is that deference is an intensional relation, and so whether you are rationally required to defer to a person at a time can depend on how that person and that time are designated. In this paper I suggest a way of completing the Reflection Principle and Group Reflection Principle, and I argue that so completed these principles are plausible. In particular, they do not fall foul of the Sleeping Beauty case (Elga in Analysis 60(2):143–147, 2000), the Cable Guy Paradox (Hajek in Analysis 65(286):112–119, 2005), Arntzenius' prisoner cases (Arntzenius in J Philos, 100(7):356–370, 2003), or the Puzzle of the Hats (Bovens and Rabinowicz in Episteme 8(3):281–300, 2011).

**Keywords** Deference · Reflection principle · Intensionality · Designator · Probability · Epistemology

## 1 Introducing the reflection principles

The original Reflection Principle (van Fraassen 1984) states—in short—that any rational agent defers to his or her future self.

✉ Anna Mahtani
A.Mahtani@lse.ac.uk

[1] Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London WC2A 2AE, UK

Reflection A rational agent defers to his or her future self

To see what 'defers to' means, take agents A and B and times $t_i$ and $t_j$, where $\mathrm{Cr}_{\mathrm{A}i}$ designates A's credence function at time $t_i$, and $\mathrm{Cr}_{\mathrm{B}j}$ designates B's credence function at time $t_j$. Then for A at $t_i$ to defer to B at $t_j$ is for the following to hold:

For any claim P and any value $v$ such that $0 \leq v \leq 1$ and $\mathrm{Cr}_{\mathrm{A}i}(\mathrm{Cr}_{\mathrm{B}j}(\mathrm{P}) = v) > 0$, $\mathrm{Cr}_{\mathrm{A}i}(\mathrm{P}/\mathrm{Cr}_{\mathrm{B}j}(\mathrm{P}) = v) = v$.

Reflection states that a rational agent defers in this sense to his or her future self.

As an illustration, suppose that I am currently wondering whether I have passed an exam (PASS). I consider the credence I will have in PASS in an hour's time. If I have a non-zero credence right now that my credence in PASS in an hour's time will be, say, 0.8, then what the Reflection Principle requires is that right now I have a particular *conditional* credence in PASS: my credence in PASS conditional on my credence in PASS in an hour's time being 0.8, will (if I am rational) be 0.8. In this way I defer to my future self: conditional on my future self having some credence in a claim, I have the very same credence in that claim.

There is a Dutch Book Argument for the Reflection Principle, but for reasons that I do not rehearse here I do not think that this argument is successful (see Mahtani 2015). Nevertheless—even without a Dutch Book Argument to motivate it—the Reflection Principle seems compelling if we restrict its application to particular sorts of cases. These are cases where the agent at the earlier time is certain that (s)he won't have 'gone wrong' by the relevant later time: (s)he won't have forgotten anything, or have become irrational, but will simply have responded rationally (i.e. by conditionalizing) on any new evidence acquired. To make this explicit, I define a relation '… respects…' as follows:

Let the set of total evidence of agent A at $t_i$ be $\mathrm{E}_{\mathrm{A}i}$, the set of total evidence of agent B at $t_j$ be $\mathrm{E}_{\mathrm{B}j}$, and $\mathrm{E}_{\mathrm{B}j} - \mathrm{E}_{\mathrm{A}i}$ be the set (which may be empty) containing any evidence that B has at $t_j$ that A does not have at $t_i$. Then A at $t_i$ respects B at $t_j$ iff A at $t_i$ is certain that B's credence function at $t_j$ is simply A's credence function at $t_i$ conditionalized on that evidence (if any) that B has at $t_j$ that A does not have at $t_i$: i.e. on $(\mathrm{E}_{\mathrm{B}j} - \mathrm{E}_{\mathrm{A}i})$.

In this definition, 'A' and '$t_i$' should be designators that A at $t_i$ would recognize as designating him- or herself and the current time respectively. To see why this clarification is needed, suppose that the designator 'A' picked out A in some obscure way, perhaps by giving the time and location of A's birth relative to the time and location of the birth of George Orwell. Then—even if A at $t_i$ somehow came to learn that B at $t_j$ has a credence function that is simply A's credence function at $t_i$ (where 'A' is an obscure designator that A cannot tell designates him- or herself) conditionalized on some additional evidence—there would be no particular reason to expect A at $t_i$ to defer to B at $t_j$. In this definition, then, we take both 'A' and '$t_i$' to be non-obscure

designators—specifically, they must be designators that A at $t_i$ would recognize as designating his or herself and the current time.

With this clarified, consider a case where agent A at time $t_0$ respects her future self at time $t_1$. Surely if she is rational A at $t_0$ will defer to that future self? After all, A at $t_0$ is certain that her credence function at that future time ($Cr_{A1}$) will differ from her current credence function ($Cr_{A0}$) only if she acquires some new (true) evidence between $t_0$ and $t_1$, in which case she will simply conditionalize on that evidence. It seems, then, that by the agent's own lights at $t_0$, $Cr_{A1}$ is an *improvement* on $Cr_{A0}$ (or at worst identical to $Cr_{A0}$), and so that A at $t_0$ should defer to her future self at $t_1$.

This suggests the following restricted version of the Reflection Principle:

Reflection* Any rational agent who respects her future self defers to that future self

By restricting the Reflection Principle in this way, many well-known counterexamples to the original Reflection Principle are disposed of. For example, one sort of counterexample to the original Reflection Principle involves an agent who suspects that she might forget something in the future (Talbott 1991); another involves an agent who suspects that she might become irrational in the future (Christensen 1991). These work as counterexamples to the original Reflection Principle because they involve rational agents who do not defer to their future selves. But they do not work as counterexamples to Reflection*: the rational agents in these cases do not defer to their future selves, but then they do not *respect* their future selves, and so Reflection* is not violated.

We can add a further restriction to the principle to avoid another sort of counterexample. In some cases, an agent's credence function is not 'transparent': (s)he is not certain what credence (s)he currently has in some claim.[1] In these sorts of cases, Reflection* can give the wrong results. Suppose for example that I am considering the claim (EXACTLY) that I have a credence of exactly 0.5 in some claim. I am unsure whether EXACTLY is true, because I don't have perfect access to my own credences: I suspect that I might have credence of exactly 0.5 in some claim, but I can't be sure. Suppose then that my current credence function is given by $Cr_0$, and my credence function a single millisecond later is given by $Cr_1$. Assume that I at $t_0$ respect my future self at a millisecond later. If I am rational, what is $Cr_0(\text{EXACT}/Cr_1(\text{EXACT}) = 0.5)$? It seems plausible to claim that my conditional credence here ought to be very high: after all, conditional on my credence in EXACT being 0.5 in a millisecond, it is very probably 0.5 right now—in which case I do have a credence of 0.5 in some claim, in which case EXACT is true. This would violate Reflection*. A more sophisticated understanding of deference should be able to handle this sort of case,[2] but for the purposes of this paper I will simply restrict the application of Reflection* to cases where all agents have the

[1] Timothy Williamson argues that conditions that are 'luminous'—i.e. 'inherently accessible to us' are very rare—and that our credences are not amongst them (Williamson 2000, p. 94). I do not dispute this: my claim is not that credences are always luminous—or that the credences of even rational agents are always luminous. My claim is just that ignorance of one's own current credences will not play any significant role in the examples discussed in this paper.

[2] For example, we might define deference as follows: for any claim P and any value $v$ such that $0 \leq v \leq 1$ and $Cr_{Ai}(Cr_{Bj}(P) = v) > 0$, $Cr_{Ai}(P/Cr_{Bj}(P/Cr_{Bj}(P) = v) = v) = v$.

relevant access to their own current credence functions: the issue that I focus on in this paper arises even with Reflection* so restricted.

Despite the adjustments that I have made above, Reflection* still faces counterexamples—and these are the 'puzzling cases': e.g. the Sleeping Beauty case (Elga 2000), the Cable Guy Paradox (Hajek 2005),[3] and Arntzenius' prisoner cases (Arntzenius 2003). These cases all involve a rational agent who respects his or her future self, and yet does not defer to that future self.[4] The difficulty is in reconciling the intuitive pull of Reflection* with the clear fact that there are cases (the puzzling cases) where Reflection* is violated. My main aim in this paper to effect this reconciliation. First though, I want to get the problem stated in its full generality, so I turn now to the Group Reflection Principle.

Reflection* focuses on the attitude that an agent should take to his or her *own future self*. But why should the principle have this narrow focus? Putting the Dutch Book Argument for the Reflection Principle aside, the motivation for accepting Reflection* is simply this intuitive thought: if you *respect* your future self, then you see your future credence function as an improvement (or at worst identical) to your current credence function—and so if you are rational you defer to that future self. This intuitive thought seems just as compelling if we drop the assumption that the respected agent must be your own future self: surely you should defer to *any* agent that you respect, regardless of who it is? This suggests a generalization of Reflection*, into this principle:

Group Reflection* For any rational agent A at $t_i$ and any agent B at $t_j$, if A at $t_i$ respects B at $t_j$, then A at $t_i$ defers to B at $t_j$

Group Reflection* is closely related to the principle that Hedden calls 'Expert Deference' (Hedden, Time-Slice Rationality 2015, pp. 23–24), and is equivalent to a version of the principle that Bovens and Rabinowicz investigate (2011, p. 293).[5] Group Reflection* is compelling for the same sorts of reasons as Reflection*: if you respect some agent-at-a-time, then you presumably see his or her credence function

[3] Hajek (2005, p. 118) does not see the Cable Guy paradox as a puzzling counterexample to the Reflection Principle, because he has accepted the restriction placed on the Reflection Principle by Schervish et al. (2004). I discuss this restriction in relation to my own position in Sect. 4.

[4] Arntzenius disagrees: he argues that in the puzzling cases that he raises, the agents involved are failing to conditionalize—even though they do not suffer from any sort of cognitive defect. On this reading, the puzzling cases are violations of Reflection [as Arntzenius writes, these 'violations of conditionalization can be parlayed into violations of reflection' (Arntzenius 2003: 370)], but they would not be violations of Reflection*, because the agent at the earlier time does not respect his or her future self (because he or she is not certain that (s)he will simply conditionalize on any evidence acquired). This would be one way to rescue Reflection* in the face of these examples. The difficulty here is that on a natural reading of Arntzenius' examples, the agents involved *are* conditionalizing.

[5] My Group Reflection* is identical to the principle that Bovens and Rabinowicz are considering when they write: 'if we modify Group-Reflection along these lines and restrict the scope of the principle to group members who have the same priors, are fully epistemically rational, and have all the evidence that we have and possibly more…' (Bovens and Rabinowicz 2011: 293). Bovens and Rabinowicz go on to argue that this principle is violated in a version of the Story of the Hats. I agree that it is so violated, though I disagree with Bovens and Rabinowicz about *why* it is violated, and about how Group Reflection* should be amended in response. I contrast my response to that of Bovens and Rabinowicz in Sect. 3.

as an improvement on (or at worst, identical to) your own—and so if rational you will defer to that agent-at-a-time.

Group Reflection* entails Reflection*: according to Group Reflection* a rational agent defers to any agent-at-a-time that (s)he respects—whether that agent-at-a-time is his or her own future self or not. Group Reflection* thus inherits all the puzzling counterexamples that face Reflection* [e.g. the Sleeping Beauty problem (Elga 2000), the Cable Guy Paradox (Hajek 2005), and Arntzenius' prisoner cases (Arntzenius 2003)]. There is also a counterexample aimed specifically at the Group Reflection Principle (Bovens and Rabinowicz 2011), and in the following section I add my own counterexample (*The Mug*) to the heap.

## 2 A problem for the reflection principles

Group Reflection* seems intuitively compelling, but a moment's consideration shows that it is incomplete. Deference is an intensional relation: whether you defer to some agent at a time can depend on how both that agent and that time are designated. Thus we must consider whether the Group Reflection* requires A at $t_i$ to defer to B at $t_j$ under *any* designator, or under just some particular sorts of designators.

First, note that just as 'deference' is intensional, so is 'respects' as I have defined it. To see this, suppose that a student Tom knows that Professor Smart is perfectly rational, shares all Tom's priors, and has all of Tom's evidence and more. Thus Tom respects Smart. But now suppose that Smart supplements his university salary by secretly working as the campus gorilla gram under the name of 'Gus'. Tom does not respect Gus: perhaps Tom thinks it possible that he, Tom, knows things that Gus does not, or perhaps Tom thinks it possible that Gus is not perfectly rational. Intuitively, Tom is rationally required to defer to Smart, but not to Gus. More generally, we should take Group Reflection* to require that if you respect an agent at a time under some way of designating that agent and time, then you are rationally required to defer to the person at that time *so designated*—but not necessarily under other ways of designating that person and time.[6]

[6] This refinement to the Reflection Principle assumes that 'referentialism about credence' (hereafter RAC) is false. Chalmers defines this principle as follows:

> I take referentialism about credence to be committed to at least the following claims… If '*a*' and '*b*' are two names for the same object, then in having a certain credence that *a* has ϕ and in having a certain credence that *b* has ϕ, the corresponding objects of credence are the same. Likewise, when one sincerely asserts '*a* has ϕ' and '*b* has ϕ', one expresses high credence in the same object of credence.
> (Chalmers 2011 p. 590)

We can see that RAC conflicts with my claim above that 'respects' and 'defers to' are intensional relations. 'Professor Smart' and 'Gus' are two names for the same object, so if RAC were true, then Tom's credence that Smart has some quality would have to equal Tom's credence that Gus had that quality. For example, if Tom is certain that Smart is rational, then Tom would also have to be certain that Gus is rational. More generally, if Tom respects Smart, Tom would also have to respect Gus. And if

Are there any further refinements that need to be made to Group Reflection*? Is A at $t_i$ rationally required to defer to B at $t_j$ under any way of designating B and $t_j$ just provided that A at $t_i$ respects B at $t_j$ so designated? In fact, this does not hold: we need to place further restrictions on how B at $t_j$ is designated to get a compelling version of the principle. To explain why we need these extra restrictions, I give a counterexample to Group Reflection* as it stands so far. This is a very clear counterexample, because if we attempt to apply Group Reflection* in this case, we reach a contradiction.

To set up the scenario, I first note that Group Reflection* can require an agent, if rational, to defer to several different people—each with different credence functions. It may be that at $t_i$ A respects both B at $t_j$ and C at $t_k$. But might this lead to incoherence? Suppose that besides having all the evidence that A has at $t_i$, B at $t_j$ and C at $t_k$ have each gained some additional evidence, but *different* evidence, resulting in B at $t_j$ having a credence of $x$ in P, while C at $t_k$ has a credence of $y$ in P, where $x \neq y$. How can A at $t_i$ defer to both B at $t_j$ and C at $t_k$ if they have different credences in P? This is easily answered: to say that A at $t_i$ defers to both B at $t_j$ and C at $t_k$ is to say that A at $t_i$ has certain *conditional* credences. From the fact that A at $t_i$ defers to B at $t_j$, it does not follow that if B at $t_j$ has a credence of $x$ in P, then A at $t_i$ has (outright) a credence of $x$ in P: rather, it follows that A's credence at $t_i$ in P *conditional* on B's having a credence at $t_j$ of $x$ in P, is $x$. Similarly, from the fact that A at $t_i$ defers to C at $t_k$, it does not follow that if C at $t_k$ has a credence of $y$ in P, then A at $t_i$ also has a credence of $y$ in P: rather, it follows that A's credence at $t_i$ in P *conditional* on C's having at $t_k$ a credence of $y$ in P, is $y$. And of course A's *conditional* credence at $t_i$ in P—conditional on two different claims (even two different *true* claims)—can take different values.[7]

We hit a problem, however, in a case where A at $t_i$ respects both B at $t_j$ and C at $t_k$, B at $t_j$ and C at $t_k$ have different credences in some claim P, *and A* at $t_i$ *knows what credences B* at $t_j$ *and C* at $t_k$ *have in P*.[8] For by Group Reflection*, A at $t_i$ is

Footnote 6 continued
Tom's credences are such that he counts as deferring to Smart, then given that Tom's credence in any claim about Smart must be the same as his credence in the equivalent claim about Gus, it will inevitably work out that Tom defers to Gus too. This gives us a reason to drop RAC, but there are other compelling and more obvious reasons to drop RAC in any case. For example, it seems obvious that Tom might have a high credence that Smart is bald (having seen him on numerous occasions), but a much lower credence that Gus is bald (having never—knowingly—seen Gus out of his gorilla costume): this is possible only if RAC is dropped.

[7] If this is not obvious, then consider that my credence that some card drawn randomly from a full pack is the ace of spades will be $^1/_{26}$ conditional on the claim that the randomly drawn card is black, and $^1/_4$ conditional on the claim that the randomly drawn card is an ace. If the card actually drawn is the ace of clubs, say, then it is true both that the card is black and that the card is an ace. Here my credence that the randomly drawn card is the ace of spades takes different values conditional on two different true claims.

[8] More generally, we would hit a problem whenever A at $t_i$ respects both B at $t_j$ and C at $t_k$, and A at $t_i$'s expectation of B at $t_j$'s credence in P does not equal A at $t_i$'s expectation of C at $t_k$'s credence in P. I focus on cases (such as my case of *The Mug*) where A is certain of both B and C's credences in P: these are just particularly clear examples of the general problem.

rationally required to defer to both B at $t_j$ and C at $t_k$. Thus A's credence at $t_i$ in P, conditional on B at $t_j$ having a credence of, say, $x$ in P, is $x$. And if A at $t_i$ knows (or, more accurately, has a credence of 1) that B's credence at $t_j$ in P is $x$, then it follows that A at $t_i$ must have an outright (i.e. non-conditional) credence of $x$ in P. Similarly, A's credence at $t_i$ in P, conditional on C at $t_k$ having a credence of $y$ in P, is $y$. If A knows at $t_i$ that C's credence at $t_k$ in P is $y$, then it follows that A at $t_i$ must have an outright (i.e. non-conditional) credence of $y$ in P. But whereas A's credence at $t_i$ in P conditional on two different claims can be different, A's *outright* (i.e. non-conditional) credence at $t_i$ in P cannot take two different values: A cannot have at $t_i$ both a credence of $x$ in P and a credence of $y$ in P. Thus if we could find a case where A at $t_i$ respects both B at $t_j$ and C at $t_k$, B at $t_j$ and C at $t_k$ have different credences in some claim P, and A at $t_i$ knows what those credences in P are—then we will have shown that Group Reflection* must be false.

Here is such a case:

The Mug[9]

You are playing a card game. There are three players—you (A), Bob (B), Carol (C)—and a dealer. One of the three players has already been randomly and secretly selected by the dealer to be the 'mug', and another to be the 'lucky player'. The dealer has dealt each player one card: the mug has been deliberately dealt a black card, the lucky player has been deliberately dealt a red card, and the other player has been dealt a card selected at random from a full pack. You and the other two players know that this is the set-up, but neither you nor the other two players know which of you is the mug, and which the lucky player. At $t_0$, the dealt cards are lying face down on the table; by $t_1$, each player will have turned over and privately looked at his or her own card. Let us assume that at $t_0$ you know that your credence function (call it $Cr_0$) is (in all relevant ways) the same as each of the other players' credence functions: you all have the same (relevant) evidence, and the same priors, and you are all rational.[10] Now consider this claim:

TWO RED Two red cards have been dealt

[9] There are a range of cases in the literature which bear similarities to my scenario here. For example, Luc Bovens and Wlodek Rabinowicz's case of the 'Story of the Hats' (Bovens and Rabinowicz 2011, p. 293); Aaron Bronfman's case of the mystery coin (Bronfman 2015, p. 1337); and John Pittard's 'Special Friend' case (Pittard 2015).

[10] It would be unrealistic to suppose that there might be three people (you, Bob and Carol) each with *exactly* the same evidence—or that you might be certain that this is so. This is why I say that you all have the same *relevant* evidence. But it is worth noting that the scenario works equally well as a counterexample to Group Reflection if we are stricter about this requirement. Suppose then that the three players live parallel, qualitatively identical lives on different planets. Each player knows that (s)he is one of these three players, but does not have any way of knowing *which* player (s)he is. The players then all have exactly the same evidence at $t_0$ (at least, exactly the same uncentred evidence), and know that this is so. Each player at $t_0$ will respect both the mug at $t_1$ and the lucky player at $t_1$. Thus we can run the example even if we are stricter about the requirement that each player should have the same evidence at $t_0$—and that you (at $t_0$) should be sure that this is so.

At $t_0$, your credence in TWO RED is $^1/_2$, and you know that each player currently also has a credence of $^1/_2$ in TWO RED. What will the players' credences in TWO RED be at $t_1$? You can already (at $t_0$) work some of this out. The mug will see that (s)he has got a black card, and so his or her credence in TWO RED will drop to $^1/_3$.[11] And the lucky player will see that (s)he has got a red card, and so his or her credence in TWO RED will increase to $^2/_3$.[12] The other player might have been dealt a red card (in which case his or her credence in TWO RED will increase to $^2/_3$), or (s)he might have been dealt a black card (in which case his or her credence in TWO RED will decrease to $^1/_3$): at $t_0$ you do not know what credence the other player will have at $t_1$ in TWO RED.

At $t_0$ you respect each player at $t_1$. For you at $t_0$ know that by $t_1$ each player will have gained some evidence—(s)he will have learnt what card (s)he has been dealt—and will have simply conditionalized on this new evidence. Thus at $t_0$ you respect both the mug at $t_1$ and the lucky player at $t_1$. Thus, by Group Reflection*, if you are rational then at $t_0$ you defer to both the mug at $t_1$ and the lucky player at $t_1$.

The problem of course is that if you are rational, then at $t_0$ you *don't* defer to the mug at $t_1$ and the lucky player at $t_1$: to defer to them both would be incoherent. To see this, consider first that if you at $t_0$ defer to the mug at $t_1$, then you at $t_0$ have a credence of $^1/_3$ in TWO RED conditional on the mug's credence at $t_1$ in TWO RED being $^1/_3$. But of course you know (have a credence of 1) at $t_0$ that the mug's credence at $t_1$ in TWO RED is $^1/_3$—so if at $t_0$ you defer to the mug at $t_1$, then you have an outright (not conditional) credence of $^1/_3$ in TWO RED. Similarly, if at $t_0$ you defer to the lucky player at $t_1$, then you at $t_0$ have a credence of $^2/_3$ in TWO RED conditional on the lucky player's credence at $t_1$ in TWO RED being $^2/_3$. But of course you know at $t_0$ that the lucky player's credence at $t_1$ in TWO RED is $^2/_3$—so if at $t_0$ you defer to the lucky player at $t_1$, then you at $t_0$ have an outright (not conditional) credence of $^2/_3$ in TWO RED. Thus if at $t_0$ you defer to both the lucky player at $t_1$ and the mug at $t_1$, then you at $t_0$ have a credence of $^1/_3$ in TWO RED and a credence of $^2/_3$ in TWO RED—which is impossible. Thus Group Reflection* is incoherent: it places inconsistent demands on a rational agent in *The Mug* scenario. I turn now to analyze why Group Reflection* fails in this scenario.

[11] To see the calculation here, let $Cr_0$ designate the players' credence at $t_0$ (given that each player's epistemic state at $t_0$ is relevantly similar), and let $Cr_{mug1}$ designate the mug's credence at $t_1$. Let's suppose that the mug happens to be Bob, so when the mug turns over his card the relevant evidence he gains is that Bob has a black card: the calculation would go through in the same way if we were to suppose instead that the mug was you or Carol. What we need to show is that $Cr_{mug1}(\text{TWO RED}) = {}^1/_3$. Thus, given that $Cr_{mug1}$ is effectively $Cr_0$ conditionalized on the claim that Bob has a black card, what we need to show is that $Cr_0(\text{TWO RED/Bob has a black card}) = {}^1/_3$. $Cr_0(\text{TWO RED/Bob has a black card}) = \frac{Cr_0(\text{TWO RED and Bob has a black card})}{Cr_0(\text{Bob has a black card})}$. $Cr_0(\text{TWO RED and Bob has a black card}) = Cr_0(\text{Bob has a black card/TWO RED}) \times Cr_0(\text{TWO RED}) = {}^1/_3 * {}^1/_2 = {}^1/_6$. $Cr_0(\text{Not-TWO RED and Bob has a black card}) = Cr_0(\text{Bob has a black card/not-TWO RED}) * Cr_0(\text{not-TWO RED}) = {}^2/_3 * {}^1/_2 = {}^1/_3$. $Cr_0(\text{Bob has a black card}) = Cr_0(\text{TWO RED and Bob has a black card}) + Cr_0(\text{not-TWO RED and Bob has a black card}) = {}^1/_6 + {}^1/_3 = {}^1/_2$. Thus $Cr_{mug1}(\text{TWO RED}) = Cr_0(\text{TWO RED/Bob has a black card}) = \frac{Cr_0(\text{TWO RED and Bob has a black card})}{Cr_0(\text{Bob has a black card})} = \frac{{}^1/_6}{{}^1/_2} = {}^1/_3$.

[12] The calculation is parallel to that in the previous footnote, except that here instead of calculating $Cr_{mug0}$(TWO RED/Bob has a black card), we calculate $Cr_{luckyplayer0}$(TWO RED/Bob has a red card).

## 3 A diagnosis of the problem, and a remedy

Viewed in a certain way, *The Mug* scenario can seem very mysterious. The mug at $t_1$ has all the relevant evidence that you have at $t_0$.[13,14] Furthermore, the mug has some additional evidence: the mug knows something that you don't, namely ($E_{Mug1} - E_0$). This additional evidence (you may be certain) leads the mug at $t_1$ to rationally reduce his or her credence in TWO RED to $^1/_3$—just as this evidence would rationally lead you at $t_0$ to do, if you but had it. Shouldn't you then follow suit and reduce your own credence in TWO RED to $^1/_3$ at once? If you know that you *would* reduce your credence in TWO RED to $^1/_3$ on learning what ($E_{Mug1} - E_0$) is—whatever that evidence ($E_{Mug1} - E_0$) turns out to be—then why not just reduce your credence in TWO RED to $^1/_3$ at once?

This sounds persuasive, but there is a jump in this reasoning which is easy to miss. It is true that if you were to learn simply ($E_{Mug1} - E_0$), then you would rationally reduce your credence in TWO RED to $^1/_3$, just as the mug has done. But it is not true that if you were to learn *what* ($E_{Mug1} - E_0$) *is*, then you would reduce your credence in TWO RED to $^1/_3$. For learning what ($E_{Mug1} - E_0$) is, involves not only gaining the evidence ($E_{Mug1} - E_0$), but also recognizing this as ($E_{Mug1} - E_0$). And if you were to learn what ($E_{Mug1} - E_0$) is in this sense—i.e. learn ($E_{Mug1} - E_0$) and also recognize it as such—then you would not decrease your credence in TWO RED from $^1/_2$: the mug was *bound* to get a black card after all.[15] Similarly if the mug at $t_1$ was to recognize the evidence he has gained as ($E_{Mug1} - E_{A0}$), then of course (s)he would increase his or her credence in TWO RED back to $^1/_2$. Thus though the mug rationally decreased his or her credence in TWO RED to $^1/_3$ on learning ($E_{Mug1} - E_{A0}$), it is just not the case that if you were to learn *what* ($E_{Mug1} - E_{A0}$) *is*, then you would do likewise. Thus the plausible-sounding argument above does not go through.

The issue here of course is that the mug does not know that (s)he is the mug. It is true that the mug has all of your evidence and more, but you need not defer to the mug on account of his or her extra evidence, because (s)he does not recognize that

[13] It is tempting to protest that there is something that you (at $t_0$) know that the mug (at $t_1$) doesn't—namely that (s)he is the mug. But what exactly is this fact that you know? If it is that the mug is the mug, then of course the mug knows this tautology too. What the mug doesn't know is that *s(he)* (you, Bob, or Carol) is the mug—but then you don't have this piece of information either. After all, *you* might actually be the mug—in which case how could you (at $t_0$) possibly know something that the mug (at $t_1$) doesn't, given that the mug does not forget anything between $t_0$ and $t_1$?

[14] The same points apply to the lucky player at $t_1$.

[15] To see this, we can calculate $Cr_0$(TWO RED/the evidence that the mug has gained (i.e. $E_{mug1} - E_0$) is that Bob has a black card). This is $\frac{Cr_0(\text{TWO RED and } E_{mug1} - E_0 \text{ is that Bob has a black card})}{Cr_0(E_{mug1} - E_0 \text{ is that Bob has a black card})}$. $Cr_0$(TWO RED and $E_{mug1} - E_0$ is that Bob has a black card) = $Cr_0(E_{mug1} - E_0$ is that Bob has a black card/TWO RED)*$Cr_0$(TWO RED) = ($^1/_3$)($^1/_2$) = $^1/_6$. $Cr_0$(not-TWO RED and $E_{mug1} - E_0$ is that Bob has a black card) = $Cr_0(E_{mug1} - E_0$ is that Bob has a black card/not-TWO RED)*$Cr_0$(not-TWO RED) = ($^1/_3$)($^1/_2$) = $^1/_6$. $Cr_0(E_{mug1} - E_0$ is that Bob has a black card) = $Cr_0$(TWO RED and $E_{mug1} - E_0$ is that Bob has a black card) + $Cr_0$(not-TWO RED and $E_{mug1} - E_0$ is that Bob has a black card) = $^1/_6 + ^1/_6 = ^1/_3$. Thus $Cr_0$(TWO RED/$E_{mug1} - E_0$ is that Bob has a black card) = $\frac{Cr_0(\text{TWO RED and } E_{mug1} - E_0 \text{ is that Bob has a black card})}{Cr_0(E_{mug1} - E_0 \text{ is that Bob has a black card})} = \frac{^1/_6}{^1/_3} = ^1/_2$.

evidence for what it is—namely the *mug's* extra evidence ($E_{Mug1} - E_{A0}$). With this in mind, we can give a new definition of 'respect', as follows:

> Let the set of total evidence of agent A at $t_i$ be written as $E_{Ai}$, the set of total evidence of agent B at $t_j$ be written as $E_{Bj}$, and $E_{Bj} - E_{Ai}$ be the set (which may be empty) containing any evidence that B has at $j$ that A does not have at $t_i$. Then A at $t_i$ respects B at $t_j$ iff A is certain that B's credence function at $t_j$ is simply A's credence function at $t_i$ conditionalized on knowledge of *what* $E_{Bj} - E_{Ai}$ *is*.

Here, knowledge of what $E_{Bj} - E_{Ai}$ is involves both knowledge of any extra evidence ($E_{Bj} - EA_i$), together with recognition that this extra evidence is $E_{Bj} - E_{Ai}$. We can see that given this new definition of 'respect', you at $t_0$ do not respect the mug at $t_1$. You are certain that the mug's credence at $t_1$ is simply your credence at $t_0$ conditionalized on the evidence that the mug has gained ($E_{Mug1} - E_0$), but of course the mug will not recognize this as the evidence that *the mug* has gained (i.e. as $E_{Mug1} - E_0$). If we now consider Group Reflection*, with 'respect' interpreted in this new way, we can see that there is no rational requirement for you at $t_0$ to defer to the mug at $t_1$—or indeed the lucky player at $t_1$, or the other player at $t_1$.[16] However, you are required to defer to yourself at $t_1$, Bob at $t_1$ and Carol at $t_1$. This is because (even under our new interpretation of 'respect') you at $t_0$ will respect yourself, Bob and Carol at $t_1$. To see this, take Bob as an example. Let the evidence that the players lack at $t_0$ and Bob has at $t_1$ be written as ($E_{Bob1} - E_0$). You can be certain at $t_0$ that Bob will know what $E_{Bob1} - E_0$ is: i.e. Bob will both have the evidence $E_{Bob1} - E_0$ and (assuming that he knows when $t_1$ comes, and knows his own name), Bob will recognize that evidence as ($E_{Bob1} - E_0$).[17] Thus you at $t_0$ are rationally required to defer to Bob at $t_1$—and similarly to yourself at $t_1$ and Carol at $t_1$: this leads to no inconsistency, however, for you do not know at $t_0$ what credences yourself, Bob and Carol will have at $t_1$.[18] Here we see again that both 'respects' and 'defers to' are intensional relations: Bob at $t_1$ and the mug at $t_1$ may be one and the same—but you respect and defer to one but not the other.

With this new definition of 'respects' in place, we can dispense with a range of potential counterexamples to Group Reflection*, including Arntzenius' Prisoner Cases (Arntzenius 2003), the Cable Guy Paradox (Hajek 2005), Sleeping Beauty

[16] Intuitively, you *should* defer to the other player at $t_1$—even though Group Reflection* (as it stands) does not require it. In this case, the other player has gained some extra evidence ($E_{Other1} - E_0$), without recognizing it as such, but that does not matter because the other player's credence in TWO RED would be unaffected by coming to recognize his or her extra evidence as *what* ($E_{Other1} - E_0$) *is*. We could adjust Group Reflection* to include these sorts of cases [and Schervish et al. incorporate this sort of refinement with their requirement 3.1 (Schervish et al. 2004, p. 317)], but for simplicity I do not do so here.

[17] Recall that Bob knows the set-up of the game, and so knows that all the players have the same (relevant) evidence at $t_0$; furthermore, we are assuming that credences are transparent in relevant respects in the cases discussed.

[18] And your credence in TWO RED at $t_0$ can consistently equal your expectation of your credence at $t_1$ in TWO RED, and equal your expectation of Bob's credence at $t_1$ in TWO RED, and equal your expectation of Carol's credence at $t_1$ in TWO RED.

(Elga 2000) and the Story of the Hats (Bovens and Rabinowicz 2011). To illustrate the strategy, below I first discuss one of Arntzenius' prisoner cases, attributed to John Collins (Arntzenius 2003, pp. 362–3), and then the Story of the Hats (Bovens and Rabinowicz 2011).

### 3.1 Arntzenius' prisoner case

At $t_0$ you are left alone in a room with two clocks, one of which (clock A) reads 6.30 p.m., and one of which (clock B) reads 7.30 p.m. You know that the two clocks run at the right rate, but you do not know which one is telling the right time: $Cr_0$ (Clock A is right) = $^1/_2$ and $Cr_0$ (Clock B is right) = $^1/_2$. You also know that a fair coin has already been tossed: you don't know the result, but you do know that if it landed heads then the light in your room will be switched off at midnight, and that if it landed tails then the light in your room will be left on all night. This scenario focuses on your credence in HEADS. At $t_0$, you have a credence of $^1/_2$ in HEADS. Should you (at $t_0$) defer to your future self at 11.30 p.m. over HEADS? After all (let's assume) you can be sure at $t_0$ that by 11.30 p.m. you will not have forgotten anything, or become irrational, but will simply have conditionalized on any evidence that you have gained. If you at $t_0$ are rationally required to defer to yourself at 11.30 p.m., then it works out that at $t_0$ your credence in HEADS ought to be $^5/_{12}$ rather than $^1/_2$. For at $t_0$ you have a credence of $^1/_2$ that by 11.30 p.m. your credence in HEADS will still be $^1/_2$ (which is what will happen if clock B is correct, and neither clock has shown midnight by 11.30 p.m.), and you have a credence of $^1/_2$ that by 11.30 p.m. your credence in HEADS will have decreased to $^1/_3$ (which is what will happen if clock A is correct, for then by 11.30 p.m. clock B will have shown midnight without the light being switched off, and so your credence in HEADS will have decreased to $^1/_3$).[19] Thus if at $t_0$ you are rationally required to defer to your future self at 11.30 p.m., your credence in HEADS at $t_0$ must equal your expectation at $t_0$ of your credence in HEADS at 11.30 p.m., which is $^5/_{12}$. Clearly, however, your credence in HEADS at $t_0$ should be $^1/_2$, so it seems that Group Reflection* gives the wrong result here.

If we use our new definition of 'respect', then the argument above does not go through. At $t_0$, you consider the evidence that you will have gained by 11.30 p.m. (i.e. $E_{11.30p.m.} - E_0$). You are certain that by 11.30 p.m. you will simply have conditionalized on this evidence. However, you will not have recognized the evidence as such—i.e. as the evidence that you have gained by 11.30 p.m. (i.e. as $E_{11.30p.m.} - E_0$). If you did recognize it as such, then you would realize that the evidence that you had gained had no significance for the likelihood of HEADS: regardless of whether the coin landed heads or tails, the light would still be on at 11.30 p.m. At $t_0$ then you do not respect your 11.30 p.m. self, and Group

[19] At $t_0$ you will give equal weight ($^1/_4$) to each of the following possibilities: Clock A is correct and HEADS; Clock A is correct and TAILS; Clock B is correct and HEADS; Clock B is correct and TAILS. If you learn that clock B has shown midnight without the light being switched off, then you can eliminate the possibility that clock B is correct and HEADS, and your credence will be divided equally among the three remaining options, giving you a credence of $^1/_3$ in HEADS.

Reflection* does not require you to defer to your 11.30 p.m. self. Again, we see that both 'respects' and 'defers to' are intensional relations—for though you do not respect or defer to your 11.30 p.m. self, you do both respect and defer to yourself at the point when clock A reads 11.30 p.m., and to yourself at the point when clock B reads 11.30 p.m.—and of course one of these selves must be identical with your 11.30 p.m. self. Deferring to each of these selves does not require you to adjust your credence at $t_0$ in HEADS from $^1/_2$.

### 3.2 The story of the hats

Bovens and Rabinowicz construct a version of a scenario they call 'The Story of the Hats', which looks like a counterexample to Group Reflection*. Bovens and Rabinowicz recognize this, and recommend a revision of Group Reflection*. Below I first describe the counterexample, and then contrast the revision to Group Reflection* made by Bovens and Rabinowicz with my own.

Here is the scenario. There are three players: Alice, Bob and Carol. At $t_0$ they are standing in the dark, and each player is given a hat which (s)he puts on. The hats are all either black or white, and each player's hat colour has been decided secretly and independently by some random process (e.g. a coin toss). At $t_1$ the lights are switched on, and each player can see the other players' hats, but not his or her own. All the players know that this is the set-up, that they are all perfectly rational, that they all share the same relevant evidence (before $t_1$) and that they all share the same priors.

Now consider this claim DIFFERENT

DIFFERENT Not all three hats are of the same colour

What credence does each player have at $t_0$ in DIFFERENT? This is easily calculated: the chance of the players all ending up with black hats is $^1/_8$, and the chance of the players all ending up with white hats is $^1/_8$, so the chance of all the players ending up with hats of the same colour is $^1/_4$. Thus at $t_0$ every player has a credence of $^3/_4$ in DIFFERENT.

At $t_1$, the lights are switched on, and each player can see the other players' hats but not his or her own. If DIFFERENT is false, then all three players see the other two players wearing hats of the same colour; if DIFFERENT is true, then just one of the players sees the other two players wearing hats of the same colour. We define the 'Selected Player' as follows: we suppose that a random ballot is held before $t_0$ to give an ordering of the players, and the players are not informed of the outcome—i.e. they are not told how they are ordered. Out of the players at $t_1$ who can see two hats of the same colour (and there will be only one such player if DIFFERENT is true), the player who came first in the ordering is the Selected Player. The Selected Player will not be able to figure out—either at $t_0$ or $t_1$—that (s)he is the Selected Player.

The Selected Player at $t_1$ will have a credence of $^1/_2$ in DIFFERENT: (s)he will see that the other two players have hats of the same colour, and will have a credence of $^1/_2$ that her own hat is that colour too. All the players are able to calculate this at $t_0$: at $t_0$ they can be certain that by $t_1$ the Selected Player will have gained some

evidence (call it $E_{S1} - E_0$), and will have rationally conditionalized on this evidence to give him or her a credence of $^1/_2$ in DIFFERENT. Should all the players at $t_0$ defer to the Selected Player at $t_1$? If so, then they ought to all have a credence of $^1/_2$ in DIFFERENT at $t_0$. But this is clearly counterintuitive: if rational, the players will have a credence of $^3/_4$ in DIFFERENT at $t_0$. Does Group Reflection* lead us astray here?

To rescue Group Reflection*, Bovens and Rabinowicz suggest the following adjustment: 'To get a tenable version of Group-Reflection, we need to restrict the principle even further and require that the group members we can rely on should have at least as much information as every other member in the group' (Bovens and Rabinowicz 2011, p. 293). The version of Group Reflection that Bovens and Rabinowicz are promoting requires an agent S at a time $t_i$ to defer to every member of group $G(S_i)$, where group $G(S_i)$ is defined as follows. First we define group $R(S_i)$ as the set of people-at-times whom S at $t_i$ considers to be epistemically rational, to share S's own priors and to have all of S at $t_i$'s evidence. Then we define group $G(S_i)$ as the set of members of $R(Si)$ who have all the evidence (and possibly more) that is available to every member of $R(Si)$ (Bovens and Rabinowicz 2011, p. 293). It is the members of $G(Si)$ that S at $t_i$ is rationally required to defer to. This Group Reflection Principle is fairly weak: it never requires you to defer to two agents who have different credence functions. And yet it is still vulnerable to a version of the Puzzle of the Hats, as follows. Suppose that the Story of the Hats is as described previously, except that now all players swallow a tablet between $t_0$ and $t_1$; the Selected Player's tablet will be a placebo, but the other players will be given a mind-altering drug; the drugged players won't notice that they have been drugged,[20] but the drug will make them irrational. All the players know in advance that this is the set-up. Now let $R(S_0)$ be the group that the Selected Player (whoever (s)he is) at $t_0$ considers to share his or her priors, and to have all his or her current evidence and possibly more, and to be epistemically rational. Who should go in this group? Certainly the Selected Player (at both $t_0$ and $t_1$) belongs in this group, but the other players should not go in the group as they are irrational.[21] The Selected Player at $t_1$ has all the evidence had by every member of $R(S_0)$, and so is a member (the only member) of $G(S_0)$. Thus on the revised Group Reflection Principle suggested by Bovens and Rabinowicz, the Selected Player at $t_0$ is rationally required to defer to the Selected Player at $t_1$. Thus given that the Selected Player at $t_0$ is certain that the Selected Player at $t_1$ has a credence of $^1/_2$ in P, the Selected Player at $t_0$ is rationally required to have a credence of $^1/_2$ in P. This is counterintuitive.

[20] This is to ensure that the selected player at $t_1$ won't immediately know that (s)he is the selected player.

[21] I said that the drug *makes* the players irrational, which might be taken to imply that at $t_0$ they are rational—but I don't intend to imply this. I am assuming that a player is a member of group $R(A_0)$ only if (s)he is *diachronically* rational, so if an agent is irrational at any time then (s)he is excluded from $R(A_0)$. An objector might claim that Alice at $t_0$, Bob at $t_0$ and Carol at $t_0$ are all rational (because they are rational at $t_0$), and so should be added to $R(A_0)$. My point will then still go through, for the Selected Player at $t_1$ has all the evidence and more than that had by Alice at $t_0$, Bob at $t_0$ and Carol at $t_0$, and so the Selected Player will still be a member (the only member) of $G(A_0)$.

My version of Group Reflection*—with the new definition of 'respect'—handles the counterexample in a different way. At $t_0$, each player can be certain that the Selected Player at $t_1$ will have gained some information (which we can designate as $E_{S1} - E_0$), and that the Selected Player at $t_1$ will simply conditionalize on this extra evidence, and so rationally decrease her credence in 'DIFFERENT' to $^1/_2$. Every player at $t_0$ would decrease his or her credence in just this way were (s)he to learn the information $E_{S1} - E_0$. However, (s)he would not decrease his or her credence in just this way were (s)he to learn *what* $E_{S1} - E_0$ *is*. For if (s)he were to learn what it is, then (s)he would know that it is the Selected Player's extra information, and of course the Selected Player was bound to see two hats of the same colour.[22] Thus under my new interpretation of 'respect', the players at $t_0$ do not respect the Selected Player at $t_1$, and so Group Reflection* does not require the players at $t_0$ to defer to him or her.

The players at $t_0$ are however rationally required to defer to this very player at $t_1$ under a different designator. The Selected Player must be either Alice, Bob or Carol, and the players at $t_0$ respect (and so according to Group Reflection* are rationally required to defer to) each of Alice, Bob and Carol at $t_1$. This points to a fundamental difference between my response and that of Bovens and Rabinowicz to this problem. The response of Bovens and Rabinowicz is to restrict the set of people (or, more accurately, the people-at-times) to whom Group Reflection* requires deference—with the result that in Story of the Hats as originally described, the players at $t_0$ are not required to defer to any of the players at $t_1$. This is not a consequence of my response: on my view, the players at $t_0$ are all required to defer to all of the players at $t_1$—but only under certain designators. Once it is acknowledged that deference is an intensional relation and Group Reflection* is interpreted with this in mind, we can see the possibility of excluding the Selected

[22] To see the calculation here, suppose that you are a player at $t_0$. Your credence in DIFFERENT is currently $^3/_4$, and our task is to calculate what your credence in DIFFERENT would be if you were to discover what it is that the selected player will learn (i.e. what ($E_{S1} - E_0$) is). Let's assume that the selected player is Carol, and that what she learns is that Alice and Bob are both wearing white hats: the calculation will run in a similar way no matter who we take as the selected player, and no matter whether we suppose (s)he sees two black hats or two white hats. Now we need to calculate your credence at $t_0$ in DIFFERENT under the condition that what the Selected Player learns ($E_{S1} - E_0$) is that Alice and Bob are both wearing white hats. First we calculate your credence at $t_0$ in the conjunction of DIFFERENT and the claim that ($E_{S1} - E_0$) is that Alice and Bob are both wearing white hats. This conjunction holds if and only if Alice and Bob are both wearing white hats, and Carol is wearing a black hat (and so is automatically the Selected Player): your credence is this outcome is $^1/_8$. Next we need to calculate your credence in the claim that ($E_{S1} - E_0$) is that Alice and Bob are both wearing white hats, and this will equal the sum of your credence in the conjunction of DIFFERENT and the claim that ($E_{S1} - E_0$) is that Alice and Bob are both wearing white hats (which we have already established is $^1/_8$), PLUS your credence in the conjunction of not-DIFFERENT and the claim that ($E_{S1} - E_0$) is that Alice and Bob are both wearing white hats. The conjunction of not-DIFFERENT and the claim that ($E_{S1} - E_0$) is that Alice and Bob are both wearing white hats obtains if and only if all three players are wearing white hats *and* Carol has happened to come out first in the random ordering (so that she is chosen as the Selected Player): your credence in this outcome is ($^1/_8$)($^1/_3$), which is ($^1/_{24}$). Thus your credence in DIFFERENT, under the condition that ($E_{S1} - E_0$) is that Alice and Bob are both wearing white hats, is $\frac{^1/_8}{^1/_8 + ^1/_{24}} = {}^3/_4$. Thus your credence in DIFFERENT at $t_0$ is $^3/_4$, and if you were to learn simply what it is that the Selected Player learns (i.e. what ($E_{S1} - E_0$) is), then your credence in DIFFERENT would be unchanged.

Player at $t_1$ from the group of people-at-times that the players at $t_0$ are rationally required to defer to, without excluding Alice, Bob or Carol at $t_1$.[23]

## 4 Reaching a consensus

It is heartening to discover that this is an area in which philosophers seem to be reaching a consensus. Closely related suggestions for refinements to the Reflection principles can be found in various forms in the literature, and below I draw out some of these connections.

### 4.1 Bronfman

Bronfman (2015) argues that 'deference to a known expert may be appropriate under some descriptions, but not others' (Bronfman 2015, p. 1333). Bronfman proposes that the 'differentiating factor is whether the expert can self-identify under the description' (Bronfman 2015, p. 1340). In this paper I have argued that both respect and deference are intensional relations—and so that whether they hold can depend on how the 'expert' is described or designated. Furthermore, on my view the conditions under which one agent respects another (designated or described in a particular way) are sensitive to whether that second agent can 'self-identify under the description'—given certain background assumptions. To see this, consider again the case of *The Mug*. Let's assume that in this scenario—as in the others that I discuss in this paper—agents have luminous access where relevant to their own

[23] The device for defining the 'selected player' does not appear in the original Story of the Hats (Bovens and Rabinowicz 2011), yet all the players at $t_0$ know that at $t_1$ there will be at least *one* player who sees two hats of the same colour, and so has a rational credence of $^1/_2$ in DIFFERENT. Are the players at $t_0$ rationally required to defer to this player, so described? If so, then it seems that my response to the puzzle will not work—for all players at $t_1$ who see two hats of the same colour will recognize this (indefinite) description as applying to themselves.

The problem with this idea is that it would require a notion of deference that could relate indefinitely described persons-at-times. Deference—as so far defined—is taken to be a relation between *specific* persons-at-times (on my view, under specific designators). Thus we know what it would be for A at $t_i$ (with credence function $\mathrm{Cr}_{\mathrm{A}i}$) to defer to B at $t_j$ (with credence function $\mathrm{Cr}_{\mathrm{B}j}$): for every claim P and value $v$ where $0 \leq v \leq 1$, $\mathrm{Cr}_{\mathrm{A}i}(\mathrm{P}/\mathrm{Cr}_{\mathrm{B}j}(\mathrm{P}) = v) = v$. And we know what it would be for A at $t_i$ to defer to *every* member of some group G: for example, if group G contains just B at $t_j$ and C at $t_k$, then A at $t_i$ defers to every member of G iff A at $t_i$ defers to both B at $t_j$ *and* C at $t_k$. We can also understand what it would be for A at $t_i$ to defer to *some* member of group G: A at $t_i$ defers to some member of group G iff A at $t_i$ defers to *either* B at $t_i$ *or* C at $t_k$. But what would it be for A at $t_i$ to defer to a member of group G, without deferring to any specific member of the group?

In the Story of the Hats, we can define group G as the set of players at $t_1$ who have seen two hats of the same colour. The players at $t_0$ all defer to every member of group G under *some* description: e.g. if Alice at $t_1$ sees two hats of the same colour, then Alice at $t_1$ is a member of group G, and of course all the players at $t_0$ defer to Alice at $t_1$—but they are not thereby required to adjust their credence in DIFFERENT from $^3/_4$. Are the players at $t_0$ also required to defer to every member of group G under that description—i.e. as *a member of group G*? Our current notion of deference does not determine what relation would be thereby required to hold between the players at $t_0$ and this underspecified person-at-a-time.

Thanks to an anonymous reviewer for this line of thought.

current credences. Thus the players at $t_0$ know what their current total evidence is, and they know that they know that this is their current total evidence. Furthermore, we select '$E_0$' to be such that the players at $t_0$ know that their current total evidence is designated by $E_0$. Thus the players at $t_0$ know what $E_0$ is. The players at $t_0$ respect the mug in the sense of 'respects' with which I began this paper: they are certain that the mug's credence function is simply their own current credence function ($Cr_0$) conditionalized on whatever evidence the mug has gained ($E_{Mug1} - E_0$). Thus the players at $t_0$ are certain that the mug at $t_1$ has any evidence that they have at $t_0$, and so given that the players at $t_0$ know what $E_0$ is (and know that they know this), they can be sure that the mug at $t_1$ knows what $E_0$ is too. Given that we are assuming luminosity holds where relevant in these examples, the mug at $t_1$ also knows what evidence (s)he has at $t_1$: if we assume further that the players at $t_0$ know that the mug at $t_1$ has luminous access to his or her relevant credences, then it follows that the players at $t_0$ know that the mug at $t_1$ knows what his or her own current total evidence is. Does it follow that the players at $t_0$ know that the mug at $t_1$ knows what $E_{Mug1} - E_0$ is, and so that the players at $t_0$ respect the mug at $t_1$—in my *new* sense of 'respect'? This does not follow: the players at $t_0$ know that the mug at $t_1$ knows what $E_0$ is, and also that the mug at $t_1$ knows what *his or her own current credence function is*, but it does not automatically follow that the mug can figure out what $E_{Mug1} - E_0$ is, because the mug may not know what $E_{Mug1}$ is. This would be guaranteed, however, if the mug at $t_1$ could 'self-identify' as the mug at $t_1$ (and if the players at $t_0$ knew that (s)he could do so), for (s)he could then identify his or her own current total evidence as $E_{mug1}$. Thus we can see how whether one agent respects another (in my new sense of 'respect') can be sensitive to whether that second agent is designated in such a way that (s)he can self-identify under that designator. Thus Bronfman and I have converged on closely related refinements to the Reflection Principle (Bronfman 2015; Mahtani 2014).

Bronfman (2015) and (Weisberg 2005, pp. 183–186) have both drawn out a connection between the requirement that an agent self-identify under some description, and the requirement that the agent's possible evidence forms a partition. The connection depends on certain background assumptions (Bronfman 2015, pp. 1346–1348), but we can get the rough idea by considering my example of *The Mug*. Take the set consisting of the pieces of evidence that the players at $t_0$ think the mug might have acquired by $t_1$ (i.e. the pieces of evidence that the players at $t_0$ think might be $E_{Mug1} - E_0$). This set does not form a partition, for the pieces of evidence are not disjoint. For example, for all the players at $t_0$ know, by $t_1$ the mug might have gained the evidence that Bob's card is black (which would be the relevant evidence that the mug acquires if the mug is Bob); alternatively, by $t_1$ the mug might have gained the evidence that Carol's card is black (which would be the relevant evidence that the mug acquires if the mug is Carol). But these two pieces of evidence are not disjoint: it might be the case both that Bob's card is black and that Carol's card is black. Thus the evidence that the mug might gain by $t_1$ (as considered from the players' position of knowledge at $t_0$) does not form a partition.

My adjustment to the definition of 'respects' ensures that whenever an agent A at $t_i$ respects (in the new sense) an agent B at $t_j$, then the set containing the pieces of extra evidence that A at $t_i$ thinks B at $t_j$ may have acquired (i.e. $E_{Bi} - E_{Ai}$), will

form a partition. This is because A at $t_i$ respects B at $t_j$ only if A at $t_i$ is certain that B at $t_j$ knows what the evidence that (s)he has acquired (i.e. $E_{Bj} - E_{Ai}$) is. Thus there can be no two *compatible* but *different* pieces of evidence $E_x$ and $E_y$ such that A at $t_i$ thinks it possible that the total extra evidence B has acquired might be $E_x$, or might be $E_y$. For A at $t_i$ will be certain that if the total extra evidence that B at $t_j$ has acquired is $E_x$, then the fact that $E_x$ is the total extra evidence that B at $t_j$ has acquired will itself be entailed by $E_x$; and if the total extra evidence that B at $t_j$ has acquired is $E_y$, then the fact that $E_y$ is the total extra evidence that B at $t_j$ has acquired will itself be entailed by $E_y$. Given that we are taking $E_x$ and $E_y$ to be different pieces of evidence, then it cannot be the case both that the total evidence that B at $t_j$ has acquired is $E_x$ and that the total evidence that B at $t_j$ has acquired is $E_y$, and so $E_x$ and $E_y$ are incompatible. More generally, if A at $t_i$ respects B at $t_j$, then the possible (from the perspective of A at $t_i$) pieces of evidence that B at $t_j$ may have acquired will be disjoint. Furthermore, given that A at $t_i$ is certain that B at $t_j$ will know what extra evidence (if any) (s)he has acquired, then (again, from the perspective of A at $t_i$) B at $t_j$'s possible pieces of evidence will be exhaustive. Thus provided that A at $t_i$ respects (in my new sense of 'respects') B at $t_j$, the set of extra evidence that A at $t_i$ thinks may have been acquired by B at $t_j$ (i.e. $E_{Bj} - E_{Ai}$) will form a partition. Several authors have shown that various versions of the Reflection and Group Reflection Principles follow automatically from the assumption that the expert's possible evidence forms a partition—given certain background assumptions (Briggs 2009; Bronfman 2015; van Fraassen 1995; Weisberg 2005).

It is interesting to note that, once again, it matters how the agents (or agents-at-a-time) are designated. The possible (from the perspective of the players at $t_0$) evidence that the mug might have acquired by $t_1$ does not form a partition, but the possible (from the perspective of the players at $t_0$) evidence that Bob might have acquired by $t_1$ does form a partition—and this holds even if Bob is in fact the mug. Of course, the evidence that an agent *actually* has does not depend on how the agent is designated: thus '… has evidence E' is not an intensional context. But the evidence that the agent could *possibly* have at a time may depend on how the agent is designated: thus '… could possibly have evidence *x*' *is* an intensional context. This is just a particular instance of the general fact that modal contexts and epistemic contexts can be intensional. Whether an agent's evidence at a time forms a partition, then, can depend on how that agent and time are designated.

### 4.2 Hedden

Hedden (2015) argues in defense of a principle that he calls 'Expert Deference', which for our purposes we can treat as identical to Group Reflection* but with 'respects' taken as defined at the start of this paper.[24] Hedden notes that Expert Deference would be incoherent if we could find a case in which it requires you to

[24] There is a difference: both principles require you to defer to any agent who you are certain has a credence function that is simply your own conditionalized on some true evidence; Group Reflection* but not Expert Deference also requires you to defer to any agent who you are certain has a credence function that is identical to your own. This difference is not important in what follows.

defer to two different agents (let's say, A and B), with different credences in P, where you know what these credences are. Hedden argues that—provided certain criteria are met—such a case can never arise. Expert Deference only requires you to defer to both A and B if you are certain that A and B are both perfectly rational, share your priors, and are certain of all that you are certain of—i.e. share all your evidence. Thus in a case where Expert Deference requires you to defer to both A and B, it must be the case that A and B are both certain that A and B share the same priors and are perfectly rational (for they are certain of all that you are certain of); furthermore, if you are certain that A has a credence of $x$ in P, and that B has a credence of $y$ in P, then you must be certain that A and B are certain of this—and so that they are certain that they are certain of this, and so on. In short, A and B would have to know each others' credences in P, know that they are both rational, know each others' priors—know that they each know all of this, and so on. Hedden then uses Aumman's 'Agreeing to Disagree' result (Aumann 1976): 'Aumann showed that if two rational agents with common priors have common knowledge of each other's credences in a proposition [P], then their credences in [P] must be same' (Hedden 2015, p. 25). Thus it seems that we simply cannot have a case where Expert Deference requires you to defer to both A and B, where A and B have different credences in P, and where you know what these credences are.

With this in mind, our case of *The Mug* may seem puzzling. Here we seem to have a case where Expert Deference requires you at $t_0$ to defer to both the mug at $t_1$, and the lucky player at $t_1$. You at $t_0$ know that the mug at $t_1$ has a credence of $^1/_3$ in TWO RED, and also that the lucky player at $t_1$ has a credence of $^2/_3$ in TWO RED. But how can this sort of case arise, given Hedden's argument? For it seems that Aumann's result should rule here that the mug at $t_1$ and the lucky player at $t_1$ both have the same credence in TWO RED. After all, they are two rational agents with common priors who have common knowledge of each other's credences in TWO RED (i.e. the mug at $t_1$ knows that the lucky player's credence at $t_1$ in TWO RED is $^2/_3$, the lucky player at $t_1$ knows that the mug's credence in TWO RED is $^1/_3$, they each know that they each have this knowledge, and so on). How then is it possible that the mug at $t_1$ and the lucky player at $t_1$ have different credences in TWO RED?

The answer is that—as Hedden observes—Aumann's result follows only given certain assumptions, and one of these assumptions is that the possible evidence that might be had by an agent at a time forms a partition. This assumption does not hold in our scenario, for neither the mug nor the lucky player's possible evidence at $t_1$ forms a partition. Thus the mug at $t_1$ and the lucky player at $t_1$ *can* 'agree to disagree'. In contrast, Bob at $t_1$ and Carol at $t_1$ cannot 'agree to disagree': the possible evidence that Carol at $t_1$ may have (from the perspective of Bob at $t_1$) does form a partition, and vice versa, and all the other relevant assumptions are met for Aumann's result to go through. But of course Bob and Carol may be the lucky player and the mug respectively, and so—intriguingly—Aumann's argument rules out two agents agreeing to disagree when designated in one way, without ruling out those same agents agreeing to disagree when designated differently.

### 4.3 Schervish et al.

Schervish et al. (2004) focus on the original Reflection Principle, according to which any rational agent defers to his or her future self. They place some restrictions on the scope of the Reflection Principle, one of which is of particular interest to us. To explain this restriction, we need the idea of a 'stopping time', and this can be explained intuitively with the following example. Suppose that I am interested in the ratio of men to women walking past my window. I sit down by the window at 10 a.m. and start gathering data, noting each time a man or a woman walks past. When should I stop gathering data, and start analyzing the data sample I have? One option is that I could stop at 11 a.m. Provided that I have a watch, this counts as a 'stopping time', because I know at every time whether or not it is 11 a.m.: in other words, I will know when to stop gathering data. Another option is that I could stop immediately after the longest continuous run of women that occurs between 10 a.m. and noon. But this is not a stopping time, because I may not know whether this time has arrived. For example, suppose that immediately before 10.30 a.m. there was a continuous run of 14 women. This is quite a long run—but will there be a still longer run if I carry on? At 10.30 a.m. I do not know.

Schervish et al. use this concept of a 'stopping time' to place a restriction on the Reflection Principle. The Reflection Principle states that an agent, if rational, will defer to him- or herself at any future time: to use the terminology of Schervish et al., the relation is between an agent *now* and the same agent *later* (with '*now*' and '*later*' designating times). Schervish et al. claim that the Reflection principle holds—provided that certain requirements are met. The one that interests us here is this: at *now*, the agent must be certain that either *later* is a stopping time (i.e. that at any given time (s)he will know whether it is *later*), or that learning at *later* that it is *later* will not affect her assessment of the relevant claim.

Though Schervish et al. do not make this point explicit, their position implies that deference is an intensional relation—and so that referentialism about credence is false. To see this, note that whether a time counts as a 'stopping time' can depend on how it is designated. For an example of this, we can return to the case where I am gathering data about the people who walk past my window. The time immediately after the longest continuous run of women that occurs between 10 a.m. and noon is not a stopping time, and 11 a.m. is a stopping time—but of course it may be that 11 a.m. *is* the time immediately after the longest continuous run of women that occurs between 10 a.m. and noon. Thus whether a time is classed as a stopping time or not depends on how it is designated. It follows then that the Reflection Principle—with the restriction from Schervish et al. in place—can rule that a rational agent defers to him- or herself at a future time when that future time is designated in one way, without ruling that (s)he defers to him- or herself at that very same future time under a different designator.

We can extend the claims of Schervish et al. to cover Group Reflection* as well as Reflection*. To do this, we introduce the idea of a 'stopping person'—where a

stopping person knows that (s)he is the person so designated.[25] Thus in our case of the mug, Bob at $t_1$ is a stopping person, but the mug at $t_1$ is not. The requirement that Group Reflection* holds only when the proposed expert is a stopping person at a stopping time, is in effect the requirement that the expert should be designated in such a way that (s)he self-identifies under that description (Bronfman 2015, p. 1430). And as I have shown, whether one agent respects (in my new sense) another, is sensitive (given certain background assumptions) to whether that second agent is designated in such a way that (s)he self-identifies under that designator: in other words, to whether the second agent is a stopping person at a stopping time.

Thus my argument is pushing in the same direction as that of Schervish et al. Besides generalizing the position, I have aimed also to provide an intuitive justification for it. Schervish et al. introduce their restriction on the Reflection Principle by relating it to the literature on stochastic processes, but it is far from clear—to a philosopher at least—what assumptions are made in this literature, how they are justified, and why and how they should be applied in the literature on the Reflection Principle.

## 5 Conclusion

The Reflection Principles have a intuitive pull. However, the principles face puzzling counterexamples where it seems as though one of these principles should operate, but then we get counterintuitive—or even inconsistent—results. In this paper I have explored the intuitive motivation behind the reflection principles, and this has led to a clarification of the conditions in which they apply. The principles are intuitively compelling only in cases where the deferrer *respects* (in my new sense) the deferred-to agent. Provided that these criteria are in place, the reflection principles do not lead us astray.

**Acknowledgments** I'd like to thank an anonymous reviewer for this journal, and audiences at the 2nd Logic and Language Conference at the Institute of Philosophy and the LSE Choice Group for very helpful feedback on this paper. I would also like to thank the Leverhulme Trust for enabling me to complete this research.



## References

Arntzenius, F. (2003). Some problems for conditionalization and reflection. *Journal of Philosophy, 100*(7), 356–370.

[25] Adam Elga can also been seen to be extending the Stopping Rule in a similar way (Elga, Reflection and Disagreement, 2007, p. 482).

Aumann, R. (1976). Agreeing to disagree. *The Annals of Statistics, 4*(6), 1236–1239.
Bovens, L., & Rabinowicz, W. (2011). Bets on hats: On Dutch books against groups, degrees of belief as betting rates, and group-reflection. *Episteme, 8*(3), 281–300.
Briggs, R. (2009). Distorted reflection. *Philosophical Review, 118*(1), 59–85.
Bronfman, A. (2015). Deference and description. *Philosophical Studies, 172*, 1333–1353.
Chalmers, D. (2011). Frege's puzzle and the objects of credence. *Mind, 120*(479), 587–635.
Christensen, D. (1991). Clever bookies and coherent beliefs. *Philosophical Review, 100*(2), 229–247.
Elga, A. (2000). Self-locating belief and the sleeping beauty problem. *Analysis, 60*(2), 143–147.
Elga, A. (2007). Reflection and disagreement. *Nous, 41*(3), 478–502.
Hajek, A. (2005). The cable guy paradox. *Analysis, 65*(286), 112–119.
Hedden, B. (2015). Time-slice rationality. *Mind, 124*(494), 449–491.
Mahtani, A. (2014). Deference and designators. Retrieved from http://www.sas.ac.uk/videos-and-podcasts/philosophy/deference-and-designators.
Mahtani, A. (2015). Dutch books, coherence and logical consistency. *Noûs, 49*(3), 522–537.
Pittard, J. (2015). When beauties disagree. In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 5). Oxford: OUP.
Schervish, M. J., Seidenfeld, T., & Kadane, J. B. (2004). Stopping to reflect. *Journal of Philosophy, 101*(6), 315–322.
Talbott, W. J. (1991). Two principles of Bayesian epistemology. *Philosophical Studies, 62*, 135–150.
Titelbaum, M. G. (2012). *Quitting certainties: A Bayesian framework modeling degrees of belief*. Oxford: OUP.
van Fraassen, B. (1984). Belief and the will. *The Journal of Philosophy, 81*(5), 235–256.
van Fraassen, B. C. (1995). Belief and the problem of ulysses and the sirens. *Philosophical Studies, 77*, 7–37.
Weisberg, J. (2005). Conditionalization, reflection, and self knowledge. *Philosophical Studies, 135*, 179–197.
Williamson, T. (2000). *Knowledge and its limits*. Oxford: OUP.