Canadian Journal of Philosophy (2023), 53: 2, 121–137
doi:10.1017/can.2023.31

ARTICLE

How I Learned to Stop Worrying and Ignore Unwelcome Epistemic Company

Adam Piovarchy

The Institute for Ethics and Society, The University of Notre Dame Australia, Sydney, New South Wales, Australia
Email: adam.piovarchy@nd.edu.au

Abstract

The problem of unwelcome epistemic company refers to the problem of encountering agreement with your beliefs from an unwelcome source, such as someone who is known to form unreliable beliefs or have values you reject. Blanchard (2023) and Levy (2023) argue that when we encounter unwelcome agreement, we may have reason to reduce our confidence in our matching beliefs. I argue that unwelcome epistemic company rarely provides reasons to reduce our confidence, and apparent successes at improving our beliefs using unwelcome company are explained by extraneous factors. Seeing why unwelcome agents are rarely evidence our belief is false requires making a distinction between two kinds of agents who regularly form false beliefs: unreliable agents and anti-reliable agents. While unreliable agents are common, they are uninformative. While anti-reliable agents would be informative, they are incredibly rare. Unwelcome agents are also rarely evidence that we have formed our own beliefs via an unreliable process, unless we have independent evidence that we are relevantly similar to them. This is hard to obtain given that unwelcome agents, by definition, have values and methods of forming beliefs that we do not find appealing. Moreover, attempts to use unwelcome company to improve our beliefs are likely to make our beliefs worse off in a number of ways. I argue we should adopt a policy of ignoring unwelcome company, letting them have little impact on our confidence in our beliefs.

Keywords: Peer disagreement; unwelcome epistemic company; unreliable agents; anti-reliable; conciliationism

> "You can always discredit one true story by setting up ten others, palpably false, parallel to it."
> H. Beam Piper, *Police Operation*

> "The world's greatest fool may say the Sun is shining, but that doesn't make it dark out."
> Robert Pirsig, *The Zen Guide to Motorcycle Maintenance*

1. Introduction

Peer disagreement occurs when I encounter someone who is in no worse an epistemic position than me, but who has arrived at a different belief than me. If we are both calculating the tip at a restaurant and arrive at different figures, I have reason to reduce my confidence in my answer. One of us must be wrong and I have no reason to suspect it is my peer, given they are in as good an epistemic position as I am.

Recently, it has been suggested there exists a mirror image of peer disagreement cases: cases of unwelcome agreement, or what we might call 'unwelcome epistemic company.'[1] The problem of unwelcome epistemic company refers to the problem of encountering agreement with your beliefs from an unwelcome source, such as someone who is known to form unreliable beliefs or have morally abhorrent values. When I encounter peer disagreement, I encounter higher-order evidence that either myself or my peer is mistaken, gaining reason to reduce my confidence in my belief due to my peer's reliability. But with unwelcome company, the other agent's *lack* of reliability, in conjunction with their agreement with me, purportedly also provides evidence of error, favouring a reduction in confidence. Consider the following examples from Blanchard (2023) and Levy (2023):

> *Lockdowns:* Jo believes *that lockdowns to control COVID-19 are unjustified*, on the basis that the economic harms lockdowns cause will result in a greater loss of life than the virus itself. But she is uncomfortably aware that this puts her in unwelcome epistemic company: the great majority of lockdown sceptics are also vocal Trump supporters, and she has no sympathy at all for their views.
> *Refugees:* Cory believes refugees commit more crime than legal immigrants or those born in the country. He's upset to discover that this puts him in the company of white nationalists.

Blanchard suggests some reasons unwelcome agents could be cause for concern:

> *Falsity*: The fact that S believes that p is evidence that $\sim p$.
> *Malfunction:* The fact that S believes that p provides evidence that you acquired your belief p via an epistemically faulty process.
> *Vice:* The fact that S believes that p provides evidence that your belief that p is connected in some way to a moral vice.
> *Implication:* The fact that S believes that p provides evidence that you missed something important about the stakes of p that is relevant to the endorsement of p.

It should be noted there are many cases where one should *not* reduce their confidence upon finding unwelcome company:

> *Blue Sky:* Shi believes that the sky is blue. He subsequently reads that Osama Bin Laden also believed that the sky is blue.
> *Hitler:* Arkady becomes a vegetarian to prevent suffering to animals. Her friend forwards her a link to a blogpost that alleges that Hitler was a vegetarian.

Why aren't these cases worrisome? Levy (2023) argues this is because there seems to be no connection between the beliefs that these unwelcome agents hold and the properties that make them unwelcome. Osama Bin Laden's beliefs about the sky are not connected to his beliefs about the merits of terrorism, for instance. Jo and Cory's company, however, plausibly believe P precisely because of their unwelcome properties.

Unwelcome epistemic company is an interesting phenomenon which has not received much philosophical attention. The aim of this paper is to provide a thorough examination of how we should think about such cases, and draw attention to a range of interacting background factors that have not yet been noticed which lead us to misunderstand when such agents can be evidence. Using Levy and Blanchard's treatments (which many readers are likely sympathetic to) as a foil, this paper argues that, in fact, unwelcome epistemic company rarely provides reasons to reduce one's confidence in a belief, and instances where it seems to have a number of factors confounding our judgments. Moreover, I argue we should develop a habit of *not* reducing confidence in our

[1] Blanchard's (2020) treatment is anticipated somewhat by Priest (2016, sec. 6).

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

beliefs upon noticing unwelcome company. We should ignore said company, because one can rarely locate which answer is true simply by moving in what seems to be the opposite direction to ideas that appear false, or which are held by the kinds of people who believe false things.[2]

This paper proceeds as follows. First, I identify a number of confounds affecting Levy and Blanchard's cases which have not been accounted for, and identify questions arising for their handling of these cases. Second, focusing on *Falsity*, I argue that unwelcome agents largely cannot be evidence for or against *P*, unless it can be shown their belief-forming processes are somehow truth-tracking. Unreliable agents, however, frequently form their beliefs orthogonally to evidence, and thus they are not a form of evidence themselves. Third, I consider whether unwelcome agents can act as evidence that we are unreliable (*Malfunction*). I argue that while this can sometimes occur, unless we have independent evidence of relevant similarity, a match won't provide new reason to change our beliefs. I close by identifying a number of ways using unwelcome epistemic agents is likely to make our beliefs worse.

2. Counterfactuals and confidence

Levy (2023) argues that our unwelcome company's belief resulting from their unwelcome properties gives us higher-order evidence that we ought to reduce our confidence in our own belief. Because our belief-formation and updating is not transparent to us—being influenced by values, priors, and traits—finding that we share a belief with someone which is counterfactually dependent on their unwelcome properties may suggest we share their unwelcome properties, which is producing our belief.

We have reason to worry when there is a "match between the content of the belief and the properties that make their company unwelcome" (101). A 'match' is worrisome because—since that belief expresses the unwelcome company's properties—noticing it in ourselves gives us reason to suspect ours is also caused by unwelcome properties in us, such as implicit bias. For Levy, "'expression' is a causal notion: the belief cannot express [Cory's] racism unless he is racist" (n2). Even if Cory is not explicitly racist, "sharing a belief that (in their case) expresses the properties that make his company unwelcome might be evidence that the belief expresses or is influenced by worrisome states of his (105)." Unwelcome company thus putatively provide us with a defeater in two ways: (i) when there are grounds to suspect our unwelcome company's belief expresses properties that make their company unwelcome; or (ii) when there are grounds to suspect our belief might have arisen via the same unreliable belief-formation process as our company. If we can rule these possibilities out, our beliefs are in the clear. If we can't, we should reduce our confidence.

It is worth briefly noting that Levy and Blanchard's position has implications for more agents than they realise. Suppose Dory believes refugees commit equal amounts of crime as nonrefugees. She notices this is "nearby" Cory's belief that they commit more crime. If Cory has reason to reduce his confidence and Dory's belief is nearby Cory's, she will have some (albeit less) reason to reduce her confidence too; factors like implicit bias might still be making her overestimate the level of crime.

I am less sure that Cory and Jo have grounds for suspicion and will start by raising some questions about Levy and Blanchard's conclusion. In particular, it is unclear how their position handles the observation that very few beliefs are not expressive of *someone's* unwelcome properties. If our beliefs can become suspect simply in virtue of someone else's unwelcome properties causing a similar belief that then imbues them with unwelcome content, we'll have to reduce confidence in

[2] A technical note: Blanchard and Levy frame their position in terms of "reducing confidence." Though it is common to think of credences running along a scale from 0 to 1, there is disagreement about whether reducing confidence should always be thought of as moving closer towards 0, increasing one's confidence that Not-*P*, or something else like "suspending judgement" (*cf.* Friedman 2013). I don't have the space to engage in such a debate, so will stipulate that my argument is conditional on a reduction in confidence that *P* necessarily entailing an increase in confidence that Not-*P*.

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

lots of our beliefs prior to actually encountering any unwelcome company. To pick just one belief, 81 million people voted for Joe Biden in 2020 (suggesting they believed he was the better candidate), and a significantly larger number of people from other countries would have if they had US voting rights. Given such large numbers, it's incredibly likely *some* were overt racists, murderers, etc. with all manner of vices, and that some formed that belief due to unwelcome properties (e.g., perhaps a white nationalist felt betrayed by Trump's policies and thought the best thing for their movement was to get him out of the spotlight, which would render their belief 'Joe Biden should be president' an expression of their white nationalism).[3] But none of this information seems like evidence that 'Joe Biden should be president' has objectionable content. Alternatively, if it is, it seems we haven't got new reason to reduce our confidence upon encountering any such people, since we could already predict someone with unwelcome properties somewhere would support Joe Biden. The set of beliefs we'll need to reduce confidence in expands even further when we start considering the counterfactual agents (and actual agents' counterfactual beliefs) that would have resulted had chancy historical events turned out differently (*cf.* Ballantyne 2014). Observations like these make it somewhat unclear where our reason to reduce our confidence is coming from.

One might think numbers matter. Perhaps it's one thing if a single unwelcome individual shares our belief for idiosyncratic reasons, but another if certain groups regularly end up having particular kinds of beliefs. But this too has odd implications: can I really work out I should favour Not-*P* simply by counting the number of unwelcome people who believe *P*? This seems to be the wrong kind of reason (whether first- or higher-order), and points to an important difference between peer disagreement and unwelcome company which counts against thinking of them as mirror images. In peer disagreement, I ought to reduce my confidence further for each additional (independently) disagreeing peer. But with unwelcome company, it isn't clear I should decrease my confidence for each additional person I meet with unwelcome properties and a matching belief. Otherwise, one should continually improve the extent to which they believe aliens don't exist with each new person they encounter at an alien conspiracy convention (a very odd way to make progress on the Fermi Paradox). Taken to the extreme, if a certain group of people believing Not-*P* is evidence for *P*, then "you're committed to thinking that *p* is less probable conditional on 100 percent of the population believing *p* than it is conditional on 75 percent of the population believing *p* and the group in question [who make up 25% of the population] believing not-*p*" (Worsnip 2023, 6).

An alternative proposal is saying we have grounds for suspicion if there is a *strong correlation* between unwelcome properties and certain beliefs.[4] The existence of many voters with welcome properties who voted for Joe Biden means there won't be an especially high correlation between unwelcome properties and believing Joe Biden should be president.[5] If other people's beliefs that *P* are almost entirely driven by unwelcome properties—evidenced by the fact that few people without unwelcome properties believe *P*—then when I find I believe *P*, there seems to be a much higher probability that I am in the group of people who have those unwelcome properties too.

Unfortunately, it will still be unclear whether unwelcome company provides a reason to reduce confidence, because anyone who knows about the correlation will possess other evidence warranting reducing confidence. To know such a correlation obtains, it is not enough to know that people with unwelcome properties believe *P* and would have believed something different if they lacked those properties. I need to also know that people who *do not* share those properties *do not* believe *P*.

[3] Even if no such voter exists, it seems somewhat precarious to hang one's argument that it is irrational to reduce one's confidence on *this* state of affairs being true.

[4] Levy mentions correlations, saying "A belief's dependence on the properties that make the company unwelcome is correlated with that belief expressing those properties." (104). But I take him to be focusing on a different correlation: the one across possible worlds or relevant counterfactuals in which the unwelcome agent holds that belief and expresses unwelcome properties, not across agents.

[5] This would also explain much of the appeal of Levy's counterfactual dependence heuristic, since this will be satisfied when unwelcome properties and (mostly) only unwelcome properties result in those beliefs.

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

And this means I need to have evidence about other agents. In short, I need to know there are other people who disagree with my belief, many of whom will be epistemic peers or epistemic superiors. (Call this 'reliable disagreement.') To the extent that unwelcome company should lead us to reduce our confidence by it being true that a strong correlation obtains between unwelcome properties and unwelcome beliefs, it will often be because we already have reliable disagreement.

This information confounds some of Levy and Blanchard's intuition pumps, which may cause readers to misidentify where our reason to reduce our confidence is coming from. In *Lockdowns*, Jo observes that "the great majority of lockdown sceptics are also vocal Trump supporters," implying there are many nonsupporters (many of whom will be reliable disagreers) who are nonsceptics. If we consider a case where we no longer know about the correlation, it becomes much less clear how individuals should believe. Consider Nnedi, who has recently arrived in the country and comes to believe that a certain drug cures COVID-19 because President Trump (who she knows little else about) said so. She discovers Trump voters have properties she finds unwelcome and share her belief about the drug. How can she know the followers would not believe Trump were it not for their unwelcome properties? She can't form this belief simply by looking at them alone. She needs some kind of contrast group (e.g., her peers in her home country) to rule out everyone else believing the same thing. Levy describes her as discovering "this belief is *predominantly* held by people who acquired it from the same source as her, and that they regard this source as reliable *only because* they hold values she regards as reprehensible' (104; emphasis added). But these details mean this is a case where Nnedi already has reliable disagreement.

3. Unreliability and evidence

The previous section raised some concerns about Levy and Blanchard's handling of unwelcome epistemic company cases. In particular, their arguments contain a clear confound (knowledge about the presence of reliable disagreers), seem to lead to unintuitive consequences (even agents whose beliefs are merely "close" to that of the unwelcome agents will have reason to reduce their confidence, and 100% of a population agreeing on P while containing N% unwelcome agents makes P less likely to be true than when 100-N% believe P), and need to explain apparent dissimilarities between peer disagreement and unwelcome company (regarding how the number of unwelcome people I encounter should affect my confidence, and the relevance of background knowledge that *someone*, somewhere is unwelcome company for unwelcome reasons).

Despite these challenges, some readers might steadfastly insist it is intuitive that agents can sometimes improve their beliefs by reducing confidence upon noticing unwelcome company, and think I have the burden of proof to demonstrate otherwise.[6] This may particularly seem the case because some of Blanchard and Levy's claims are qualified in ways that make them seem quite modest and thus hard to challenge. Blanchard argues for "the possibility of a problem" (538), while Levy argues our company's belief "might" have arisen in a way that gives us "a reason to suspect" our own (105).

But such claims, strictly speaking, could have been established by identifying a single case where unwelcome company gives us reason to reduce our confidence. I take it that this is not all Blanchard and Levy wish to establish, as this would not show (emphasis all added) that unwelcome epistemic company "is a problem of *everyday* life" (Blanchard 2023, 530), that when confronted "this possibility *should not be dismissed* without ruling out that one is not guilty by epistemic association" (538), and that any match between our belief and our company's belief "*is sufficient* for the company to provide higher-order evidence against a belief" (Levy 2023, 104). Rather, the reasons provided are

[6] Though I would remind them that I am something of an epistemic peer who has reached a different belief to them on this matter…

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

supposed to be strong enough to "commend various forms of doubting, revising, and rechecking one's beliefs or character" (Blanchard 2020, 536).[7]

Retreating to the claims that there is only "the *possibility* of a problem" (Blanchard 2020, 538) and that "Double-checking *can* be rational" (533) would reduce the contribution of these arguments, and fail to establish that cases like *Refugees* (which are rather common and of interest to many politically engaged agents) are useful illustrations of how unwelcome company gives us reason to reduce our confidence. I take it that part of our interest in this topic comes from our desire to develop heuristics or priors with which to reason, or to identify circumstances in which we genuinely ought to change our beliefs.

In any case, I will now argue that any apparent intuitiveness is misleading because further confounds affect the cases Levy and Blanchard describe (and other work on peer disagreement). These confounds aren't merely additional variables provoking a wayward intuition. Rather, they are the result of misunderstanding how particular claims map onto states of the world featuring other agents. For the moment, let's focus on *Falsity*, and we will return to Levy's argument emphasising *Malfunction* or *Vice* later.

Let's also make a much-needed distinction between two things we might mean regarding 'unreliability,' which commonly clouds thought experiments featuring agents who are more likely to be wrong than right. Let's say that when someone's answer has little or no correlation with the correct answer, they are 'unreliable.' Someone who averages around 50% when guessing the result of a fair coin toss is unreliable regarding what side the coin will land on. Contrast this with someone whose answer is consistently the opposite of the correct answer (or strongly negatively correlated with the answer), who we'll call '*anti*-reliable.' Someone whose continued coin flip guesses average 0% would be perfectly anti-reliable.[8]

Where we draw the line between unreliability and anti-reliability depends on the nature of the task and our priors. When predicting the result of a fair dice roll, someone who gets the answer wrong 5/6ths of the time is only unreliable, not anti-reliable. This can be seen by realising that someone who gets the answer right 2/6ths of the time would be very useful to have around when betting, despite getting the answer wrong most of the time. One would do much better following this person's predictions than they otherwise would.

Likewise, finding someone who is anti-reliable within a certain domain would be quite useful. Knowing that someone is anti-reliable at picking whether a stock will go up or down tomorrow is equally as useful as knowing that someone can reliably pick whether said stocks will go up or down. Thus a principle like 'believe the opposite of what anti-reliable people believe' is potentially as rational a strategy as 'defer to reliable people's beliefs.' Let's say that when making predictions which act as evidence for us, a person whose coin-flip guesses average 0% and someone whose guesses average 100% are equally *informative.* An anti-reliable person who averages 0% is *more* informative than a reliable person who averages 90%. That anti-reliability is informative also means that someone can become *more* reliable (in the sense of being less anti-reliable) and yet *less* informative (because they also become more unreliable).

Against this backdrop, noticing unwelcome company intuitively seems like cause to reduce one's confidence. After all, it is easy to imagine or remember observing agents who regularly form their

[7] This exegetical detour is needed, as Levy and Blanchard's various descriptions lead some anonymous reviewers of this article at another journal to think my challenges were attacking a straw man. A further dialectical hurdle, which careful readers should note, is that Blanchard's position is qualified in such a way that makes it difficult for any possible objection to receive uptake: if unwelcome company "*always* provides *some* defeasible reason" (538, emphasis added), then no matter what counterexample is given, or what confounding variable is identified, one can reply the reason was either defeated, or remains but is incredibly weak. Presumably, there is such a thing as being so weak we would be better off rounding down to zero than attempting to incorporate this into our credences (cf. whether observing a non-black non-raven is *very weak* evidence that all ravens are black).

[8] Such a distinction is anticipated by Priest (2016, sec. 6), though she doesn't note the limits on the informativeness of anti-reliable agents (281).

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

beliefs in ways the evidence does not recommend. However, this reasoning moves too quickly, and the intuitive usefulness of principles like ‘reduce confidence in any belief held by an anti-reliable agent’ risks misattribution of the successes it apparently generates. This happens in two ways: first, we forget that believing the opposite of what someone believes often involves selecting a much greater set of options and, second, we misunderstand the extent to which agents who continually form false beliefs can be informative.

To see the first point, consider predicting what dice face will be rolled. Suppose it is an open question whether Peter is anti-reliable. Can one do better than Peter by reducing their confidence in a particular option that Peter chooses? Of course; if he believes it will land on 3, one will do better than he by believing ‘Not-3.’ But this is because ‘Not-3’ is a broader answer than ‘3.’ One can do better than Peter using this strategy *even if* Peter is not anti-reliable at all, i.e., if he is as equally unreliable as you would be. One might be able to do better with this strategy than Peter even if he is more reliable than average. Thus it is important to ensure that when we consider the plausibility of various approaches to unwelcome company, any success from decreasing our confidence in *P* is not simply from the fact that Not-*P* encompassed a larger range of options. We need to ensure any apparent successes in our believing are due to the anti-reliable agents in particular.

Still, reducing one’s confidence in whatever an anti-reliable agent believes does help us do better, even after correcting for an expanded range of acceptable answers. Suppose now that Peter is, in fact, somewhat anti-reliable such that if he predicts 3, this is good evidence the alternative options are each more than 1/6 likely to come up. Knowing he is anti-reliable does help us do better than we otherwise would. But it’s important to note that, absent any other evidence, the degree to which we increase our confidence in Not-3 needs to be spread across *all* other options we are considering within Not-3. While knowing what Peter believes is helpful, this is rarely as helpful as having a reliable agent—call him Steve—who gives us evidence about where to concentrate our expected probability.[9] In order for Peter’s belief to be more informative than Steve’s, Steve’s confidence needs to be significantly weaker than Peter’s, and certainly cannot be more than 20% for any particular number. If Peter is certain the dice will come up 3, such that we know it will not come up 3, we have 1/6th probability to assign elsewhere, and will likely take each other option to now have a 1/5th probability of being correct.[10] But if Steve is more than 20% confident it will come up 4 and historically well-calibrated, we will do even better by ignoring Peter altogether and simply following Steve. Knowing that Reliable Steve predicts 4 with confidence *C* is much more informative than knowing that Anti-reliable Peter predicts 3 with confidence *C*. Because there are so many more ways to be wrong than to be right on most questions, knowing that one option is wrong doesn’t especially help us identify the right answer.

This is not to say that anti-reliable agents are necessarily less informative than reliable agents. One notable instance in which they are more informative concerns Russian Roulette–style cases in which there are many more ways one can be right than wrong. However, it is clear that the majority of questions we are uncertain about are not like this. Nevertheless, one might think *Falsity* has been vindicated: if our unwelcome agent is anti-reliable, knowing what they believe gives us reason to increase our confidence in Not-*P*. Although in many cases Not-*P* will encompass an infinity of options and so not be particularly useful, in other cases we may have narrowed down the possibilities to a smaller set of options, and in *these* cases, our unwelcome agent will be particularly informative.

[9] Can we do even better still by *combining* their answers? *Only* if Peter’s informativeness is *independent* of Steve’s. If Peter achieves his informativeness simply by asking what Steve thinks and usually doing the opposite, we’ll end up overconfident taking Peter’s answer to be additional evidence after already taking into account Steve. As we’ll see, it is incredibly rare for anti-reliable people to achieve informativeness independently of reliable agents.

[10] Although Cory would improve his beliefs by assigning more confidence to refugees committing somewhat less crime, without further evidence he will also assign more confidence to them committing equal crime, no crime, and even decreasing the crime rate, all of which are false.

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

But notice that there is an inherent tension in this thought. For this to work, we need to attribute to unwelcome agents the reliability needed to arrive at a set of options *we* considered plausible, and then only after this, within that set, suddenly attribute to them the power to be anti-reliable in a way that will be informative to us. If the white nationalists believed that refugees commit 'all the crime' which was never under consideration by us, ruling that option out doesn't give us any extra probability to concentrate elsewhere. This tension is not yet fatal to Levy and Blanchard's position, but as we shall see, it results from some additional, unrecognised hurdles that unwelcome agents face before they can be informative.

4. Anti-reliability and rarity

Let me now draw attention to an underappreciated consequence of the fact that different kinds of questions can be asked in different ways. Levy and Blanchard (and much of the peer disagreement literature) tend to focus on questions that are formulated in true/false terms, e.g., "the lockdowns cause more harm than they prevent." Such framing can be useful, as it makes it easier to think of success in terms of percentages, and to imagine everyone falling on a scale from "highly reliable" to "highly unreliable." Sometimes we are required to come down clearly on one particular side of an issue, as when we vote. And in such instances, someone who was anti-reliable *would* be very informative.

But such people are going to be *much* rarer than readers may realise. Consider an otherwise ordinary clueless agent, who knows very little about the world. Ignorance is very easy to notice when asking open-ended questions such as "Who is the current president?" where there are many possible answers. When given a set of open questions, completely clueless agents score close to 0%. But once we frame our question in true/false terms, completely clueless people no longer score close to zero. Unless we've managed to pick questions that commonly mislead people, completely clueless agents instead average close to 50%. Consistently getting the wrong answer on a set of yes/no questions is *much* harder than on a set of open questions.

This is another means by which many of us overestimate the informativeness of people who routinely get the wrong answer. We imagine that someone who knows very little, and who would score close to 0% on open questions, would also score close to 0% on forced-choice questions, but they would not. People who routinely get things wrong provide very little information about which is the correct answer and thus what to believe because there are so many different ways in which one can have a wrong answer. When we provide questions that can only have one correct answer or one incorrect answer, someone who is anti-reliable would be informative. But the overwhelming majority of agents who we have previously identified to be unreliable, or unwelcome, *are not anti-reliable* in this fashion. Unless someone has observed a set of agents historically scoring worse than chance (e.g., less than 50% on true/false questions), we should be extremely hesitant to take seriously any claims that a set of agents qualifies as generally anti-reliable. Far more common is that we observed them scoring close to 0% on open questions and confused the informativeness of unreliable and anti-reliable agents.

Many readers are, at this point, thinking of counterexamples to my position. It seems unchallenging to imagine agents who we could very plausibly observe being historically anti-reliable, even while avoiding the numerous confounds I have identified thus far. My contention is that such cases will almost always involve unidentified reliable agents doing the work. Consider:

> *Security:* Charles is a security guard at a university and attends many talks by scientists as part of his job. Because he thinks the scientists are part of a global conspiracy, he believes the opposite of what they do. You do not know much about scientists and universities (or cannot access said talks), but you know Charles and his habits. You adopt a policy of believing the opposite of any belief he formed in opposition to these scientists.

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

This is a case in which "believe the opposite of what an anti-reliable person believes" is a rational strategy. However, some things need to be noted. The first is how rare it is for us to have access to an anti-reliable person that has access to a reliable source that we cannot access. This case is not at all representative of most encounters with unwelcome agents. The second is that the rationality of using Charles only gets off the ground because Charles's anti-reliability is achieved by piggybacking off the scientists' reliability. Not believing what the anti-reliable agent believes is simply an indirect way of getting at the reliable people.[11]

Consider also:

> *Supreme Court:* US President Donald Trump is picking a Supreme Court judge. You, an Australian, know relatively little about the candidates he could pick from. But being left-leaning, you know that whoever he picks, this will be evidence that *that* candidate is (by your lights) a poor candidate.

This seems to be a particularly plausible case of an informative, anti-reliable agent, from which other cases could easily generalise. However, some care is needed: Is Trump *actually* best described as anti-reliable? There is a sense in which he qualifies (for you, on this question). But it is important to note that if we instead asked "Which candidate is likely to repeal Roe vs. Wade?," "Which candidate will make rulings that Republicans like?," or any number of other questions bearing on who is best, Trump is probably going to give very similar answers to both left-leaning politicians and nonpartisan experts. That's to say, the rationality of reducing our confidence in whatever he picks only gets off the ground by attributing to Trump *epistemic reliability.* It's in virtue of having observed him historically make other similar choices that *reliably* align with his values that we're confident *this* choice will be bad by your lights, *not* his history of (say) making rudimentary errors or inconsistent assertions. If one insists on calling him anti-reliable, one runs the risk of incorrectly implying that he would give wrong answers to these other, importantly related questions. For this reason, I think it is worth distinguishing anti-reliable agents (who would also give the wrong answer to questions like "Which candidate will repeal Roe vs. Wade?") from agents that are reliable with different values (who would not).[12]

Having acknowledged that agents can be informatively anti-reliable when there is some unrecognised reliable agent in play, my contention is that—given existing physical laws— it is generally implausible that we will observe anti-reliable agents once we rule these cases out.[13] Note that being *systematically* anti-reliable requires one be *systematically responding* to evidence bearing on various questions in a particular way, namely, in the opposite direction to what said evidence recommends. But systemically responding to evidence *just is* the challenge that agents face when attempting to be reliable. We've already seen how scoring 0% on iterated coin tosses is equally as difficult as scoring 100%. But on most other tasks, anti-reliable agents necessarily face more difficulties than reliable agents, which can only decrease what we might think of as the 'transmission fidelity' between their evidence and their beliefs. Remember here that as such agents become more reliable and thus less anti-reliable, they become more *un*reliable and thus *less* informative.

An important source of difficulty is that one cannot—in ordinary circumstances—intend to be consistently anti-reliable in their belief-forming. It is frequently beyond our capacity to do so, as the literature on doxastic involuntarism has noted (e.g., Alston 1988; Chuard and Southwood 2009). No matter the incentive, I am simply unable to currently believe by will that the US is still a colony of

[11] In case one is tempted to think that their side of politics is much more likely to be correct than the other such that using political opponents would be a useful strategy, see Joshi (2022) for objections.

[12] Cases like this plausibly account for intuitions favouring *Implication*: if one cannot ask these other questions, one can infer something about their answer from who Trump thinks is best. In case readers are tempted to insist Trump could qualify as morally anti-reliable, further objections below will apply.

[13] Setting aside fanciful thought experiments featuring, e.g., magical genies, evil demons.

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

Great Britain, for instance. Admittedly, I could indirectly cause myself to form some false beliefs (e.g., signing up for brainwashing). But even an agent who is committed to being anti-reliable for some perverse reason requires some minimal degree of reliability within certain domains to navigate the world, which we have a habit of bumping up against. And it would be extremely difficult to predict ahead of time which facts an agent has the most internal reasons to form reliable beliefs regarding, and on which topics the agents' motives will better be served by forming anti-reliable beliefs. These extra difficulties mean that, at least at a certain level of generality, and given a particular set of evidence, there can never be agents who are more anti-reliable than otherwise similar agents are reliable.

Additional hurdles are in store. Return to Cory, who believes that refugees commit more crime than nonrefugees. Suppose both that refugees in fact commit somewhat less crime than nonrefugees and that we have observed white nationalists be historically anti-reliable on questions involving social groups. Our dice example showed that we are prone to overestimating how much Cory can improve his beliefs by decreasing his confidence, since we forget that he will increase his confidence in a number of other options that are also incorrect (e.g., equal crime, substantially less crime, no crime). But a greater source of concern is that Cory is at significant risk of having his beliefs end up worse off simply by having the question under consideration framed differently. If Cory is instead asked "Do refugees commit an *equal* amount of crime to nonrefugees?" he will observe the unwelcome agents believe "no," which matches his belief, and then incorrectly reduce his confidence in this belief!

This observation gets us to the heart of why agents cannot be systematically anti-reliable: it would make your beliefs incoherent. For the white nationalists to be anti-reliable in any useful sense, they would need to be disposed to give the wrong answer to "Do refugees commit more crime than nonrefugees?" as well as "Do refugees commit the same amount of crime as nonrefugees?" and other related questions. But such anti-reliability would require their beliefs to change depending on how we ask the question, and it is simply implausible this could occur.

Readers might grant this point, but still insist that unwelcome agents could be anti-reliable on some narrow questions. They might be tempted to reply that surely we can work out *which* questions the white nationalists will be informative on, and thus frame the question accordingly. But notice what this requires. In addition to requiring enough evidence to establish they are anti-reliable (not merely unreliable), *and* requiring that the white nationalists be reliable enough to narrow down *their* options to those *we* were considering, only to then exhibit anti-reliability within that set, we now *also* need it to be the case that we somehow possess enough evidence to discern how to frame our question in such a way that avoids them giving us the correct answer *without* us already possessing enough evidence to render their answers uninformative given our priors. But if we have *that* amount of evidence, I submit that it is incredibly unlikely that the white nationalists' beliefs will be informative: we've probably already got our credences where they should be. It is precisely when we have multiple options under consideration that anti-reliable agents are useful because they help us eliminate contenders. Having enough evidence to distinguish questions leading to informatively incorrect answers ("Do refugees commit more crime than nonrefugees?") from questions that will give correct answers which we should not decrease our confidence in ("Do refugees commit roughly the same amount of crime as nonrefugees?") means we probably no longer need to ask unwelcome agents what they believe.[14]

Taking stock: there are a number of reasons why it seems we can improve our beliefs using unwelcome company, but these are largely independent of the unwelcome agents or their

[14] Another exception where unwelcome agents can be informative concerns the reliability of third-party agents. Suppose I know that members of group *X* are biased towards incorrect beliefs about lockdowns. Suppose I encounter someone not of group *X* who is against lockdowns and who, given my priors, acts as new evidence lockdowns are bad. If I can find evidence they are *relevantly similar* to group *X* (beyond the bare fact of sharing *X*'s belief) then I will have reason to think that the same bias of group *X* is operative in this person, and decrease the weight I give to their testimony.

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

unwelcome properties. Thought experiments feature a number of confounds, such as knowledge of reliable agents' beliefs. Reducing one's confidence in *P* upon noticing unwelcome company often leads us to implicitly select a much wider range of options. Anti-reliable people seem common because we commonly encounter people who do not know the answer to easy questions. But we mistakenly imagine that people who score 0% on open questions will also score 0% on forced-choice questions, leading us to overestimate how informative unwelcome agents are. We confuse people whose beliefs are formed orthogonally to the evidence with people whose beliefs are consistently the opposite of the correct answer, which would require them to be truth-tracking. Once these types of questions are distinguished, anti-reliable people either provide very little information and are almost indistinguishable from merely clueless or unreliable people (because they've been asked open questions with a very large set of possible answers, and excluding one wrong answer does little to tell us which is the right answer) or it is incredibly unlikely that we could observe such people in an informative way (because being anti-reliable on forced-choice binary questions would require systematically responding to the evidence just like reliable people do, while also overcoming additional hurdles). To the extent that there are informative anti-reliable people, it's often because their informativeness piggybacks on reliable people who we can access ourselves. When we rule these cases out, being consistently anti-reliable would make the agent incoherent. Even if we restrict our attention to narrow questions, the evidence that allows us to know on which questions the agents will be anti-reliable is also evidence likely to make their answer uninformative.

5. Are we unreliable?

Having shown we should not worry about *Falsity*, let's now return to Levy's argument that unwelcome agents can be evidence our beliefs express unwelcome properties or result from an unreliable process. Let me acknowledge there are some cases where Levy's prescription seems correct. For example:

> *Depression:* You notice you believe none of your friends like you. You know this is the kind of belief often formed by agents who have depression, which is an unreliable process. You infer you should reduce your confidence in no one liking you until you have ruled out having depression.

Levy's diagnosis works well here: my belief matches that of known unreliable agents, so my belief may have resulted from a similar process and I should reduce my confidence until I have checked my reasoning. The trouble is this kind of reasoning also seems to lead to cases like this:

> *Apple:* You notice you believe the apple in front of you is red. However, you know some people falsely believe items are red because they have failed to notice a red light is present. Your belief matches the belief of these unreliable agents whose belief results from a flawed process. You should reduce your confidence until you have checked there are no red lights around.

Levy acknowledges we do not have reason to reduce our confidence when we are "entitled" to our beliefs. But the defeaters he is worried about are hard to detect and apply potentially very broadly: implicit bias (in ourselves or others; *cf.* Machery 2022), other kinds of biases (Wikipedia currently lists over two hundred), and misleading evidential environments. It is not hard to create variants of *Apple* that involve the possibility of hidden red lights, fake apples, apple-shaped pomegranates, or philosopher friends playing a prank, which by stipulation are all difficult to detect, and which we also would not be entitled to ignore as possibilities.

If we want to avoid falling into generalised scepticism, we need to understand why it seems we ought to reduce our confidence in *Depression* but not *Apple*. My contention is that in the former, we know not only that that belief can result from an unreliable process, but that that process is

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

prevalent *in groups of agents I am a member of.* In *Depression*, what gives me reason to reduce my confidence is knowing that depression is not uncommon among humans like me, and that this process often leads to that kind of belief. If I have prior reason to think I am not in this group (say, I know I have positive affect and feelings of self-worth), then I will not have reason to check: among agents like me, *that* unreliable process is rare.

The trouble for Levy and Blanchard is that unwelcome agents are, by definition, *relevantly different* to us given we don't share their values or approve of their reasoning, and this entitles us to quite a bit of confidence that some kinds of unreliable processes are not affecting us. Isn't the matching belief evidence of relevant similarity? In the way that thinking an apple looks red is evidence of being relevantly similar to apple-observing agents under red lights, yes. But this doesn't amount to much without other evidence to estimate how common that unreliable process is among a set of agents we are a member of.[15] Given I historically haven't spent much time around photo labs with red lights or fake apples, I shouldn't worry.

A crucial observation about Levy and Blanchard's treatments of individual cases is that because they consider multiple potential defeaters, some slippage occurs between the reference groups and thus belief-forming processes under consideration. If the unreliable process is 'explicit racial prejudice,' this is the type of factor Cory can introspectively be confident he does not hold and thus does not need to worry might be at play upon noticing his unwelcome company. But when the potential defeater under consideration is 'implicit bias,' the reference group of relevant unwelcome agents *changes.* Cory should no longer be thinking about 'white nationalists' but 'agents affected by implicit bias.' Whether Cory is entitled to his belief depends on how entitled he is to think his belief wasn't produced by implicit bias, which in turn depends on his independent evidence of not only how prevalent implicit bias is among agents like him, but how likely a belief like his is to result from such a process compared to a reliable process. Importantly, the white nationalists (who have explicit bias) are not good evidence of either of these things.[16]

In thinking about these cases, we also need to be careful to not confuse fortuitous prompts that cause us to remember other evidence we possess and revise our confidence, and evidence that we ought to revise our confidence. In peer disagreement, finding out what my peer believes is surprising, and this new information gives me reason to reduce my confidence. But once we begin thinking about the prevalence of unreliable processes, particularly those that are more commonplace, the unwelcome agents themselves will often drop out of the picture. When Cory changes to considering how probable it is his belief was produced by implicit bias rather than a reliable process, he is likely to answer this by referring to other evidence—e.g., statistics, journal articles, testimony from experts—much of which he probably has as background beliefs. Individual unwelcome agents count as evidence if he had only just learned about this kind of unreliable process, and in that he could, in principle, try to gather a sample of agents who share his belief, and then assess what proportion of them have their belief caused by implicit bias. But this is unlikely to be the means by which he reassesses his beliefs. Even in *Depression*, most agents in that position don't *learn* about depressed agents and then reduce their confidence. They *already* have some rough sense of the prevalence of depression and have simply been *reminded* such an unreliable process could be at play.[17] Note agents could be prompted by nonevidence (e.g., a picture of W. K. Clifford) to undergo similar self-reflection for any belief they have: they might realise that believing everyone likes them

[15] If it seems difficult to discern whether we are relevantly similar because the process under consideration is difficult to detect, we need to examine its prevalence among groups we know we are members of, e.g., "humans" or "adult Americans."

[16] Similarly, when Jo thinks lockdowns are unjustified, the relevant reference group is going to be much broader than "Trump supporters." Because she is worried about overvaluing the economy, and she consciously finds the supporters' values unwelcome, the relevant reference group will be "agents who *unknowingly* overvalue the economy." This also is the kind of feature whose prevalence the supporters are not clearly evidence of.

[17] Strictly speaking, a perfect reasoner won't reduce their confidence because the probability they are depressed will already be factored into their prior for believing that no one likes them. This is a significant contrast to cases of peer disagreement, where

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

could be produced by arrogance, believing they are more likeable than average could be produced by the better-than-average effect, believing they have average likeability could be caused by the middle option bias, etc.

To bring this all home: what gives an agent reason to reduce their confidence is *learning* they have incorrectly assessed the *prevalence* of an unreliable process *among agents like them.* All three components are needed. If we learn about an unreliable process in agents like us but have no estimate of prevalence, then we are merely sceptics listing possible defeaters (e.g., hidden red lights). If we learn an unreliable process is very common in agents completely different to us, *we* have no cause for worry: we're not them. That most drunk people overestimate their attractiveness gives me no reason to doubt my sober self-assessments. And if an unreliable process is present among agents like us but we have already taken this into account, observing others' beliefs result from said process won't give us reason to change our confidence unless they somehow show our estimate was off.[18] Given we are not perfect updaters, it will often be the case that we haven't taken all our evidence into account, but the bare fact that unwelcome agents have unwelcome properties causing a matching belief doesn't show this, especially if we lack other evidence that we are like them.

Levy is correct that unwelcome agents give us reason to reduce our confidence when noticing their belief is a means by which we learn about an unreliable process in agents like us, or learn our estimated prevalence was off. But for this to occur in practice, a *lot* turns on what our priors are (particularly if we are relevantly different to the unwelcome agents in question), what other evidence we already possess (a significant amount of which we must already have given we can identify that other agents have unwelcome properties and this is producing their beliefs), and demonstrating that unwelcome agents aren't merely fortuitous reminders of such evidence. He gestures towards some unreliable processes that are plausibly affecting both Jo, Cory, and the unwelcome groups, such as media bias. But what matters is showing show how the matching beliefs make those processes more likely to be affecting them than would have already been estimated given the agents' other evidence of such processes. In the absence of this detail, agents like Cory need not worry much about *Malfunction* or *Vice.* If one still finds it unintuitive to think Cory could arrive at such a belief and not have reason to reduce his confidence, here is a plausible explanation for why: either you possess evidence he lacks, or he possesses evidence you lack.

one cannot have high credences that the other agent is both a perfect epistemic peer and will disagree; hence disagreement is surprising. Obviously we are not perfect updaters, but such idealisations bring into focus what things are evidence compared to what are merely prompts causing us to remember other evidence. In Levy's terms, our grounds for suspecting our belief arose by the same process, or expresses the same properties, will, in many cases, be what's giving us reason to reduce our confidence independent of the unwelcome company themselves. This difference may seem like hair-splitting, but it is important to note since if one's prior is where it should be (because they're aware of those grounds) before noticing the unwelcome agent, reducing confidence upon noticing the unwelcome agent will necessarily make one do worse.

In theory, if our priors are where they should be, we should find the beliefs of the unwelcome agents *un*surprising. That we are surprised can be evidence that we had underestimated the prevalence of that unreliable process, but it is *also* evidence that the agent is not unwelcome. (Compare: finding out an expert disagrees with *P*, which you took to be well-established, is both evidence that Not-*P and* that they are not an expert.) Although it is tempting to think we can antecedently rule out the latter (and thus should take surprise to be evidence of the former), the evidence which would justify holding our knowledge of their unwelcome properties producing their belief fixed is often the type of evidence that would render the match unsurprising. Insofar as we do experience surprise in practice, the unwelcome agents are likely acting as a mere prompt.

[18] To answer some earlier questions: the number of unwelcome agents who believe *P* does give us reason to reduce confidence insofar as they provide reason to *revise* our estimate of the prevalence of that process in agents like us. If we already have an accurate estimate of that prevalence, observations of individual unwelcome agents will not matter. If there are relevant differences between us and such agents, then we also do not need to worry, which explains why the bare fact that some idiosyncratic unwelcome agents share our belief is not cause for concern either.

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

6. 'Reversed stupidity is not intelligence'

I've argued there are many considerations we need to keep in mind when trying to use unwelcome company to improve our beliefs, and there are multiple ways we overestimate our successes. But since I agree there are some cases in which unwelcome company gives us reason to reduce our confidence, and that Levy's diagnosis is strictly correct, some readers may have the impression that there is a problem of unwelcome company, it is just less frequent than it initially seemed.

I think something important would be missed by taking there to remain a problem of unwelcome company, but for it to just apply in rarer circumstances than previously suggested. I think there would be a significant mistake of emphasis. I take it that Blanchard and Levy think this is a problem that is sufficiently frequent, serious, or noticeable to be worth having on our radar, or worth drawing attention to in an academic journal, especially given the apparent parallels between such cases and peer disagreement. As already noted, the case used to first motivate the problem of unwelcome company—*Refugees*—is not an uncommon topic of discussion, many people find sentiments like Blanchard's intuitive, and Blanchard (2020) thinks the possibility of there being some malfunction in our reasoning is "especially plausible" (532). In contrast, I've argued there are important differences between unwelcome company and peer disagreement, and that quite a lot of hurdles need to be overcome before reducing our confidence is warranted.

I've mentioned that part of our interest in this topic stems from wanting heuristics or priors with which to reason, or to identify circumstances in which we genuinely ought to change the way we are believing. Given such interests, I think Blanchard and Levy have effectively focused on what should be the exception rather than the rule. Rather than getting into a back-and-forth over whether, in some additional circumstances, we can potentially squeeze out slightly more reason to reduce our confidence, so one can claim that unwelcome agents are technically evidence, we would do well to instead consider what that rule might be.

I would like to consider a phrase used by AI alignment theorist Yudkowsky (2015), and develop it into a more explicit principle:

> '*Reversed Stupidity Is Not Intelligence*' (RSINI): One can rarely, if ever, improve their epistemic position simply by doing the opposite of people who hold beliefs we find immoral or unreliable.

By "doing," I mean a range of epistemic activities such as believing, trusting, inferring, and increasing or reducing confidence in. Upon noticing that someone's belief (or an idea, or claim, or assertion) seems "stupid" (or immoral, or unjustified, or formed in an unreliable way, or too hasty, etc.), one cannot (or, at best, can only very rarely) significantly improve their own epistemic position by changing their own beliefs (increase/decrease confidence, infer, trust) in what seems to be the opposite direction.

The strength of this principle is not only that it helps keep our attention on relevant evidence and correct perceptions that reducing our confidence upon noticing unwelcome company has been a useful strategy in the past. It's that making a habit of following this principle will actively prevent us from making a variety of mistakes which people like Jo and Cory will not otherwise realise they have made. To see what I mean, let's consider a variety of ways in which attempts to improve one's beliefs using unwelcome company can go wrong, and which a full accounting of Levy and Blanchard's recommendations must keep in mind.

One class of cases are where unwelcome properties cause unwelcome agents to become *more* reliable in their belief-forming. For example, a criticisable trait like an inflated sense of superiority may in fact result in said company being highly motivated to do good research, or being scrupulous with which sources they use to inform their beliefs (e.g., maybe they only admit information from prestigious sources, and perhaps there is a correlation between prestige and veracity in some domains). Admittedly, a reliable agent with the same amount of post-investigation evidence might form the same belief. But if we do not have access to said evidence, cannot know what investigations

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

the agent has done, or take the unwelcome properties to show we cannot trust them to have conducted their investigation well, we will find it hard to know what this person would have believed if we removed their unwelcome property while holding all other features fixed.

Another class are instances where we misidentify who counts as unwelcome company or are mistaken about what a supposedly unwelcome group believes. Levy is sensitive to how we might end up with unwelcome beliefs due to the media presenting biased evidence, but we also need to consider the risk that the media has mislead us about who is unreliable and in what ways, e.g., by only presenting that group in a negative light or the most extreme examples of said group. For example, people often have mistaken beliefs about what their political opponents believe (Levendusky and Malhotra 2016). We are also prone to 'the ultimate attribution error' (a relative of the more well-known 'fundamental attribution error') where we overattribute the shortcomings of out-groups to representative, stable features of their character (Pettigrew 2020).

Alternatively, *we* might be the ones with unwelcome properties like bias. Noticing that other people don't share our values or beliefs, we might think they are unwelcome and thus unreliable. Misidentification is especially likely to happen if one has both relatively demanding values and an idiosyncratic set of beliefs. One might, for example, believe that almost all of society has been arranged to mistreat them and their in-group, taking nearly everyone else's everyday actions or beliefs to be immoral. As a result, everyone is required to conform to a set of demanding and very narrowly prescribed actions, values, and beliefs in order to actively confront this maltreatment. Someone who accepts this worldview will notice there seems to a strong correlation between holding unwelcome properties (since almost everyone else has different values and attitudes) and many unwelcome beliefs (since this group's beliefs are so idiosyncratic that few nonmembers hold them) that makes said belief seem caused by their unwelcome properties. This, in turn, means that in many instances where one finds themselves with a belief shared by most people but not shared by fellow adherents, they will take themselves to have reason to decrease their confidence in that belief. Despite gaining more and more evidence that they should reduce their confidence in their current worldview due to reliable disagreement, they will instead interpret this as reason to reduce their confidence in any beliefs incongruous with said worldview.

One final risk of using unwelcome company is that we sometimes take certain beliefs to be *constitutive* of unwelcomeness, which will result in us discounting genuine evidence. Some beliefs are genuinely constitutive of unwelcome company, such as 'Group *x* deserves less rights than group *y*.' But it is easy for people to take other beliefs such as 'refugees commit more/less crime than nonrefugees' (that are clearly historically contingent and require some empirical investigation to assess) to themselves be inherently constitutive of being unwelcome company. This may, in turn, lead to nearly any evidence favouring that belief also becoming suspect ('you can't trust those statistics'). Rather than noticing that many people who hold welcome properties hold the relevant belief and taking this to show that reducing confidence is not recommended, agents might instead take their previously welcome company to be now unwelcome. At least, it will be much harder to give uptake to evidence that doesn't come from agents the hearer already thinks are reliable.

Such risks seem especially likely in contexts of political polarisation or within echo chambers. With political polarisation, there are clearly delineated camps which make it easy to consider agents unwelcome and for it to seem like their beliefs are counterfactually dependent on the properties that make them unwelcome (since if they didn't have their unwelcome properties, they'd be in our camp with our beliefs). In an echo chamber, where dissenting views are systematically excluded and discredited (Nguyen 2020), one continually accumulates more evidence favouring their views as any disconfirming evidence is not given uptake (since it never gets admitted in the first place, or does get admitted but gets discredited). As a result, one is unlikely to notice or give uptake to reliable information coming from unwelcome company, they are likely to misidentify the extent to which certain unwelcome properties correlated with certain beliefs in others, they will accumulate more and more evidence that endorsing certain beliefs are themselves good evidence of unwelcomeness, and if they attempt to question whether such beliefs should be discounted they risk being excluded

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

or undermined themselves. When in a politicised environment or an echo chamber, adopting a heuristic of discounting beliefs that match the beliefs of unwelcome company thus seems likely to only exacerbate the extent of polarisation and strength of the bubble.[19]

In contrast, *Reversed Stupidity is Not Intelligence* avoids all these traps. Agents who follow RSINI can gain the benefits of believing welcome information produced by unwelcome agents. Even if they are no less prone to forming misleading stereotypes and committing the fundamental attribution error in judgements about other agents, these factors won't then go on to reduce the quality of their beliefs as much as they would if agents took said stereotypes to be accurate representations of unwelcomeness. They will find it much easier to give uptake to information incongruous with a demanding and idiosyncratic ideology. They will not fall prey to the trap of considering an ever-expanding set of agents to be unwelcome. And even if one suspects that a certain field or institution has biases that may warrant distrust, one will do much better, epistemically speaking, by trying to counter said biases or finding independent evidence to weigh, rather than adopting a knee-jerk reaction in the opposite direction to anything such institutions claim.[20]

Following this principle also has epistemic value outside of noticing we have unwelcome company. For example, it reminds us of the importance of engaging with the strongest arguments available for positions, where possible. Since even well-supported positions can attract bad arguers with weak evidence, RSINI reminds us that we should not take our refutations of bad arguments to be evidence that the strong arguments have been refuted.

RSINI also focuses our attention on ensuring our beliefs are formed for the right reasons in a robust manner. For example, in using guilt-by-association (e.g., 'But *S* believed *P*'), one implies that if we find out unwelcome agents hadn't believed *P* ('Hitler was a vegetarian'), our reasons to believe Not-*P* would be weaker. But there seems to be something undesirably precarious about having the beliefs we end up with being affected by what unreliable people happen to believe rather than the evidence that bears on *P* itself.

Finally, RSINI also acts as a counterweight to other habits we might have that lead us to miss out on epistemic goods. For example, some people may, at various points, have been reluctant to take the arguments of either atheists or Christian apologists seriously because they were reluctant to either affiliate (or to be perceived as affiliating) with such parties (e.g., Richard Dawkins). The point here isn't that said people made good arguments, or even that one should or should not be an atheist. It is simply that a fact such as 'Richard Dawkins [who you may think has unwelcome properties] believes God does not exist [for reasons traceable to his putatively unwelcome properties]' is not the right kind of evidence in favour of believing God does exist. One can improve their epistemic position by noticing when their mental habits are being led by a desire to not affiliate rather than evidence and reason, and RSINI can help one notice and correct for such habits.

6. Conclusion

Finding one's self with unwelcome company can feel concerning and may generate an impulse to immediately revise one's beliefs in a direction opposing one's company. However, many instances

[19] Conversely, noticing that one has beliefs shared by unwelcome company is pro tanto evidence one has not succumbed to group-think or isn't in an echo chamber with the company they consider welcome (see also Joshi (2022) for a similar line of thought). These risks also highlight some costs of politicising epistemically important institutions, such as scientific research or academia. Even if one thinks it is impossible to be truly apolitical, openly letting such institutions espouse values that large numbers of agents do not share, when combined with a susceptibility to reducing confidence in beliefs held by company that appears unwelcome, can cause large numbers of people to dismiss strong evidence.

[20] A potential objection here is that most people are bad at doing their own research. This may be true compared to simply deferring to reliable sources, but this is a case where the reliability of said sources is not recognised. What matters is that it is not clear that agents would do any better by not trying their own research and continuing to believe the opposite of the reliable sources they consider to be suspect.

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press

in which our company seems to generate reasons to revise our beliefs are often provided by other factors. There are many more ways to be wrong than to be right, unreliable people are common but not all that informative, and there are many ways in which trying to believe the opposite of what unwelcome agents believe can lead us astray. When we notice these considerations rather than insisting on successfully finding the narrow set of cases in which our impulse is justified, we should instead be more inclined to question and ignore our impulse. In some instances, we can improve our beliefs by reducing our confidence in beliefs shared with unwelcome company, but these are unlike the majority of cases that typically cause people concern. Reversed stupidity is not intelligence, and one can rarely get closer to truth simply by opposing beliefs held by agents one finds unwelcome.

Acknowledgments. For helpful comments and discussion, I am very grateful to Mark Colyvan, Nevin Climenhaga, Michael Nielsen, Mark Alfano, Sam Shpall, Shang Long Yeo, Timothy Scriven, and Neil Levy. The opening quotes and some examples were originally used by Eliezer Yudkowsky, who declined an offer to be listed as co-author.

Adam Piovarchy is a research fellow in the Institute for Ethics and Society at the University of Notre Dame, Australia. His research focuses on moral responsibility, blame, and hypocrisy.

References

Alston, William P. 1988. "The Deontological Conception of Epistemic Justification." *Philosophical Perspectives* 2: 257–99.

Ballantyne, Nathan. 2014. "Counterfactual Philosophers." *Philosophy and Phenomenological Research* 88 (2): 368–87.

Blanchard, Joshua. 2023. "The Problem of Unwelcome Epistemic Company." *Episteme*: 20 (3): 529–541.

Chuard, Philippe, and Nicholas Southwood. 2009. "Epistemic Norms without Voluntary Control." *Noûs* 43 (4): 599–632.

Friedman, Jane. 2013. "Suspended Judgment." *Philosophical Studies* 162 (2): 165–81.

Joshi, Hrishikesh. 2020. "What Are the Chances You're Right about Everything? An Epistemic Challenge for Modern Partisanship." *Politics, Philosophy and Economics* 19 (1): 36–61.

Joshi, Hrishikesh. 2022. "The Epistemic Significance of Social Pressure." *Canadian Journal of Philosophy* 52 (4): 1–15.

Levendusky, Matthew S., and Neil Malhotra. 2016. "(Mis) perceptions of Partisan Polarization in the American Public." *Public Opinion Quarterly* 80 (S1): 378–91.

Levy, Neil. 2023. "When Is Company Unwelcome?" *Episteme* 20 (1): 101–6.

Machery, Edouard. 2022. "Anomalies in Implicit Attitudes Research." *Wiley Interdisciplinary Reviews: Cognitive Science* 13 (1): e1569.

Nguyen, C. Thi. 2020. "Echo chambers and epistemic bubbles." *Episteme* 17 (2): 141–61.

Pettigrew, Thomas F. 2020. "Intergroup Attribution." In *Oxford Research Encyclopedia of Psychology.*

Priest, Maura. 2016. "Inferior Disagreement." *Acta Analytica* 31 (3): 263–83.

Worsnip, Alex. 2023. "Compromising with the Uncompromising: Political Disagreement under Asymmetric Compliance." *Journal of Political Philosophy* 31 (3): 337–57.

Yudkowsky, Eliezer. 2015. *Rationality: From AI to Zombies.* Berkeley: Machine Intelligence Research Institute.

Cite this article: Piovarchy, A. 2023. How I Learned to Stop Worrying and Ignore Unwelcome Epistemic Company. *Canadian Journal of Philosophy* 53: 121–137, doi:10.1017/can.2023.31

https://doi.org/10.1017/can.2023.31 Published online by Cambridge University Press