Rationality and Higher-order Awareness

Sturgeon, Scott

DOI:
10.1163/18756735-00000157

BRILL

GRAZER PHILOSOPHISCHE STUDIEN 99 (2022) 78–98

Grazer Philosophische Studien
brill.com/gps

Rationality and Higher-order Awareness

Scott Sturgeon
University of Birmingham, Birmingham, UK
s.sturgeon@bham.ac.uk

Abstract

It is argued that higher-order awareness is central to one type of everyday rationality. The author starts by specifying the target notion of rationality, contrasting it with other useful notions in the neighbourhood. It is then shown that the target notion relies on first-person awareness of the unfolding of cognition. This is used to explain the kernel of truth in epistemic conservatism, the structure of defeasibility, and the root motive behind the widely accepted distinction between rational inference and trivial entailment.

Keywords

defeasibility – entailment – higher-order awareness – inference – rationality

1 The Target Rationality

A major concern in the theory of rationality is how opinion should be shifted upon receipt of new information. It is argued that the answer to this question centrally involves an agent's capacity to deploy something like a *de se* running tab of their own mental states. To a rough first approximation: the thesis defended is that forward-in-time rationality is fixed at least in part by an agent's first-person capacity to track backward-in-time rationality. As we'll see, the relevant tracking is central to rationality even when it concerns only first-order matters of fact.

We begin our discussion with a pair of distinctions. Both are entirely common-sense, but only one, the first considered here, has a name in the literature. To see it, suppose Sherlock Holmes and his side-kick Watson jointly investigate a murder. During the investigation evidence crops up to establish

that the butler is guilty. Holmes and Watson both believe that the butler is guilty. Watson has visceral bias against Scots, however, and the butler happens to be Scottish. This is why Watson believes that the butler is guilty. Holmes believes on the basis of evidence unearthed in the investigation. This makes for a difference in rationality of the two beliefs that the butler is guilty, despite their sharing a content, and despite their both existing within a univocal evidential setting. Both beliefs are rational in light of the evidence, by stipulation, but only Holmes' belief is anchored to the evidence in the right way. In the vernacular of modern epistemology: both beliefs are "propositionally" rational, but only Holmes' belief is "doxastically" rational.[1]

To a rough first approximation: propositional rationality requires quality evidence to hand, and doxastic rationality requires belief to be based on quality evidence to hand. Since Holmes and Watson have good evidence that the butler is guilty, their respective beliefs are propositionally rational. Since Holmes' belief is based on good evidence, his belief is doxastically rational as well. Since Watson's belief is based on bigotry, his belief is not doxastically rational. The distinction is basically that between having good evidence and believing for good evidence.

Our second distinction is best understood by analogy. Suppose you meet someone, Happy, who happens to maximize happiness through her action. Whatever Happy does creates maximal happiness in relation to the range of options before her. Let us suppose that this occurs by chance – or perhaps a benevolent demon – and Happy has no idea it is the case. If we stipulate that producing happiness is a morally good thing, it follows that Happy's actions deliver the goods. Hence those actions deserve credit for delivering the goods. But Happy herself does not deserve any credit. It is a good question why she fails to merit credit for the moral goods delivered by her actions. It is not a good question whether she fails to deserve credit for that. It is a common-sense fact that Happy does not deserve credit for the moral goodness delivered by her actions. Common-sense recognizes the potential, then, for a disconnect between the moral status of an agent and the moral status of her actions.

Now suppose you meet someone, Verity, who has only true beliefs about the future. Suppose a benevolent demon manipulates the world to ensure this is so. The head-to-world correlation guaranteed has little to do with Verity. If we stipulate that having true beliefs is an epistemic good thing, it follows that Verity's beliefs about the future deliver the goods. But it does not follow that Verity herself deserves credit for delivering those goods. It is a common-sense

1 Roderick Firth initially drew the common-sense distinction in (Firth, 1978). For influential discussion see Kvanvig and Menzel (1990).

fact that she merits next-to-no credit for the epistemic goods delivered by her beliefs about the future. It is a good question why this is so, but it is not a good question whether Verity fails to merit credit in a case like this. Common-sense likewise recognizes the potential for a disconnect between the truth-based epistemic status of a believer and the truth-based epistemic status of her beliefs.

This sort of disconnect can also occur when epistemic status turns on evidence rather than truth. Just think of the famous movie inspector Clouseau. After each investigation, he ends-up with mountains of evidence for whatever he happens to believe about the subject-matter investigated. But Clouseau never believes in line with his evidence because of investigative diligence. It is always a fluke that his beliefs on the job are propositionally rational. He deserves no credit for that rationality. Intuitively put: his beliefs deserve credit for having contents which are backed-up by good evidence to hand, but Clouseau deserves no credit for that being the case.

From an epistemic point of view it is a good thing when beliefs manifest propositional rationality. We mustn't assume, though, that an agent deserves credit for having propositionally rational beliefs. If her cognitive efforts qua epistemic agent helped to bring about the alignment of content and evidence, the agent deserves credit for propositional rationality. If they do not do so, the agent deserves no credit. The crucial thing is whether the agent herself plays the right kind of role, *qua* agent, in the production of propositional rationality. If such rationality results from epistemic agency, credit for that rationality washes over to the agent in question. If propositional rationality springs from factors having little to do with an epistemic agent *qua* agent – as it does with Inspector Clouseau – the agent deserves no credit for having propositionally rational beliefs.

Our second distinction is thus between agential and non-agential rationality. The former occurs when rationality of belief is due – to a significant extent, at least – to the epistemic agency of the believer. The latter occurs when the rationality of belief is not due – to any significant extent, at least – to that agency. When rationality of belief is agential, a believer deserves credit for rationality of belief. When the rationality of belief is non-agential, a believer deserves no such credit. The distinction is really between situations in which the epistemic status of belief washes over to that of the agent, and situations in which the epistemic status of belief does not wash over to the rationality of the agent.

We have two common-sense distinctions before us: one separates propositional and doxastic rationality, the other separates agential and non-agential rationality. Since the distinctions are compatible on their face, there are four notional possibilities in the neighbourhood:

- Propositional agential rationality springs from quality evidence had by an agent thanks to their epistemic efforts qua agent.
- Propositional non-agential rationality springs from quality evidence had by an agent little-to-no thanks to their epistemic efforts qua agent.
- Doxastic agential rationality springs from an agent basing belief on quality evidence to hand, where basing is an exercise of epistemic agency by the agent.
- Doxastic non-agential rationality springs from the basing of belief on quality evidence to hand, where basing is not an exercise of epistemic agency.

Since each of these four notions has a diachronic and a synchronic variety, we're really faced with an eight-fold array of phenomena. The story we're about to tell could be told for each of the eight notions, but that would make for an indigestible discussion. So we restrict our attention to one of the notions in play, namely, doxastic agential diachronic rationality. This is the sort of rationality that plays out over time and springs from an agent meriting credit for the exercise of epistemic agency via the basing of belief on quality evidence to hand. This is the target notion of our discussion. From now on – unless explicitly stated otherwise – this will be the target notion of rationality in play.

2 Reasons and Defeaters

A natural way to think about evidence-based agential rationality is via the thought that people exploit reasons and defeaters. To see how this works, consider three propositions:

Heads = the claim that a coin landed heads on a particular fair toss.
Big-Tex = the claim that Texas is bigger than Germany.
Liberal = the claim that Baxter is a liberal Arkansan.

Suppose you watch the relevant coin being tossed and for no reason at all come to believe Heads. Suppose you came to believe Big-Tex for a specific reason in the past, but you've lost touch with that reason entirely. Since you come from a proud Texas family, though, which discusses all things Texas *ad nauseum*, and vets its discussion with Google, you continue to believe Big-Tex despite having lost touch with the specific reason for which you came to believe it in the past.[2] Suppose finally that you believe Liberal because you know Baxter was raised in

2 The Big-Tex example is modelled on the India case in (Christensen 1994, 74).

Eureka Springs, Arkansas: a place from your youth, a place with strong liberal traditions, pride in progressive politics, a well-known left-wing oasis in the sea of conservatism that is Arkansas.

Intuitively, then, your Big-Tex belief and your Liberal belief are rational and your Heads belief is irrational. A natural explanation for this is that your Big-Tex belief and your Liberal belief are based on reasons, while your Heads belief is not based on reasons. Intuition about rationality seems shaped by the idea that it springs from the exploitation of reasons.

There is an important asymmetry, though, in the reasons used for belief in Liberal and those used for belief in Big-Tex. The former support Liberal specifically, while the latter do not support Big-Tex in that way. Your reasons for Liberal make the claim probable on your evidence, or likely to be true from your point of view, or something like that. They are "content-specific", as we might put it. But your reasons for Big-Tex do not support Big-Tex in a similar way. You are aware that you come from a proud Texas family, and so on; but none of that supports the particular idea that Texas is bigger than Germany. Big-Tex is the hypothesis that Texas is bigger than Germany. So none of the story about your family supports Big-Tex as such. Instead, your family story counts as a reason to believe that you are something like a conditional indicator of Big-Tex's truth-value. The reasons underneath your (continued) belief in Big-Tex form into a specific reason to believe that Big-Tex is true if you believe it, and false if you reject it. They do not form into a hugely *strong* reason to believe those conditional claims, of course, but they do support the idea that your take on Big-Tex (if you have one) indicates its truth-value.

Let's mark this asymmetry by saying that you have a content-specific reason to believe Liberal and an environmental reason to believe Big-Tex. Your reason to believe Liberal is specific to its content, after all, and your reason to believe Big-Tex has mostly to do with creating a friendly intellectual environment for respecting your take on Big-Tex. To repeat: this is not because your family story makes for a specific reason to think Big-Tex is true. It is rather because that story forms into a specific reason to believe that your take on Big-Tex (if you have one) is an indication of its truth-value: if you believe it, it is likely to be true; if you reject it, it is likely to be false.

We have two kinds of reasons here: those indicating a specific claim is true and those which make for a friendly intellectual environment. The distinction is relative, of course. The belief that you come from a proud Texas family etc. is an environmental reason to believe Big-Tex, but it is a specific reason to believe that your family is aware of its background, that your family is not from New York, and so forth. When we speak of the exploitation of reasons in what follows, we'll mean the apt use of specific and environmental reasons.

In addition to being produced by good reason, rational belief can be undone by it as well. When this happens we say that rationality is "defeated" by new information. The basic idea is that old information can make for rationality and then new information can wipe it out. And it turns out there are two general ways new information can do this. Each corresponds to a type of reason just seen – specific reason and environmental reason. Consider each type of reason in turn:

A. Sometimes a person can start out believing P rationally but come to have new specific reason for P. So long as the new specific reason is strong enough, the person is no longer situated rationally to believe P. The original rationality has been defeated by new information. In a moment we'll see how this can happen when the initial rationality springs from a specific reason for P or an environmental reason for P (or both).

B. Sometimes a person can start out believing P rationally on the basis of the view that condition C indicates P's truth – which view may be held implicitly or explicitly – but then come to possess news which makes clear that C *fails to indicate* P's truth. The person is then no longer well placed to believe P rationally. Their original rationality has been defeated. As we'll see this too can happen when the initial belief is based on a specific reason for P or an environmental reason for P (or both).

Call the first sort of defeater a "negation defeater" and the second sort of defeater an "indicator defeater". Negation defeaters do their work by supporting the negation of a defeated belief's content. Indicator defeaters do their work by attacking something the agent takes to indicate the truth of the defeated belief's content.[3]

We have two kinds of reason and two kinds of defeater. This makes for a quadruple of fundamental ways that rationality can be undone: when specific reason runs into negation defeater, when specific reason runs into indicator defeater, when environmental reason runs into negation defeater, and when environmental reason runs into indicator defeater. Since a single belief can be supported by both kinds of reason, and both types of defeater can attack that support, the base cases mix in various ways. We'll restrict our attention to the base cases, however, and leave combinations of them to the reader's imagination.

3 Negation defeaters are also known as "rebutting" defeaters, and they're traditionally contrasted with "undercutting" defeaters. But the latter are a more restrictive category than our indicator defeater, as we'll see. John Pollock introduced undercutting defeat in (Pollock, 1967). His mature views on the topic can be found in (Pollock, 1987).

Recall your belief in Liberal – the claim that Baxter is a liberal Arkansan – and your belief in Big-Tex – the claim that Texas is bigger than Germany. Recall also that you have specific reasons to believe Liberal (to do with Baxter's home town) and environmental reasons to believe Big-Tex (to do with your big Texas family). The rationality of each belief can be defeated by new information in the two ways just mentioned. There are four base cases to consider:

(i) Suppose Baxter reveals in discussion that it is of first importance to him to be different than those around him. You learn that Baxter has a visceral need to stand out, and nothing bothers him more than political group-think. For this reason, Baxter supports Fox News, Donald Trump, and QAnon. This gives you strong specific reason to believe that Baxter is not a liberal Arkansan. Your initial specific reasons to believe Liberal are defeated by powerful negation defeaters. You rationally believed Liberal on the basis of specific reasons, then came to possess stronger specific reasons to believe the negation of what was initially supported by your evidence. You thereby end-up poorly placed to believe Liberal after receipt of new information. Initial rationality produced by specific reasons was wiped out by negation defeaters.

(ii) Suppose you recently visited Eureka Springs for the first time in decades. The town has changed dramatically since your last visit. It is no longer dominated by art shops, quirky restaurants, cool places with a hippy history. Now Eureka has McDonalds, Walmart, and Chick-fil-A, along with its more traditional lefty venues. You discover the population has changed in line with the shops. Politics in Eureka is now evenly split: 40% conservative, 40% liberal, 20% don't-care. Once you learn this your rational belief in Liberal is defeated by new information. That information is not specifically reason to believe not-Liberal, but it is specific reason to believe that growing up in Eureka is no longer an indicator of left-wing politics. Your new information suggests that your old specific reasons for accepting Liberal do not indicate its truth. We have the second kind of defeat mentioned before: you rationally believe Liberal on the basis of specific reasons, then come to have specific reasons to think initial reasons do not indicate Liberal's truth. Rationality is wiped out by indicator defeater.

(iii) Suppose you know the world's expert on states and countries, someone who has spent decades investigating which state is bigger than which, which country is bigger than which, and so on. You ask them if Texas is bigger than Germany. To your surprise the expert insists that Texas is not bigger than Germany, and that the view that it is is itself nothing more than an urban myth. In the event, the expert's testimony is specific reason

to believe that Big-Tex is not true. Your new information defeats earlier environmental reasons to believe Big-Tex. Here we have rationality built on environmental reason wiped out by negation defeater.

(iv) Suppose you learn that your family hate Birkenstock shoes. In fact they hate them so much that they downplay everything to do with Germany. They wish to minimize the influence of any country so corrupt as to produce Birkenstock shoes. For this reason, your family under-describe how often Germany has won the European Cup, how many people live in the country, how big the country is geographically, and so forth, and they never fact-check their discussion of Germany. This new information gives you a specific reason to think that your old reasons to believe Big-Tex do not actually create a friendly intellectual environment for that claim. After all, Big-Tex is about Germany as much as it is about Texas, and your family discussions of Germany are every bit as unhinged as its discussions of Texas are fact-checked. For this reason, you are not well-placed to regard your take on Big-Tex as a conditional indicator of its truth-value. Your environmental reason to do with Big-Tex have been defeated by an indicator defeater.

When it comes to exploiting reasons and defeaters, then, there are specific reasons to believe, environmental reasons to believe, negation defeaters for both kinds of reason, and indicator defeaters for both kinds of reason. Rationality springs from specific and environmental reasons, and it's undone by negation and environmental defeaters.

3 Basing of Belief on Reasons

Rational capacities are exercised in a great many ways. This makes the nature of inquiry itself a complex and fascinating topic.[4] We focus here on a particular aspect of rational capacity: namely, the exploitation of reasons to believe. This involves the basing of mental states on one another across time. To see this, consider three ordinary situations. The first is

The Fruit Case

Suppose you rationally believe that there are apples or oranges in the fridge, and you also rationally believe that there are no apples in the fridge. Should you come to believe that there are oranges in the fridge?

4 See Friedman (2020).

A natural thought is that it depends on how you got into your initial epistemic position. If you came to believe that there are apples or oranges in the fridge solely on the basis of rational belief that there are apples in the fridge, for example – via trivial logical inference – but then rationally reversed your take on apples for new reason, then, in that case, you should not go forward in thought with your apples-or-oranges belief to conclude that there are oranges in the fridge. Your intellectual history requires, after sensibly reversing your take on the apples, that you drop the apples-or-oranges belief and think solely that there are no apples in the fridge. On the other hand, if you start out rationally believing that there are apples or oranges in the fridge on the basis of your partner having told you as much – after he put up the shopping – then you should go forward with your apples-or-oranges belief to conclude that there are oranges in the fridge. Rationality in the Fruit Case looks to depend on how mental states are based on one another in the recent past.[5] This is at least partly why there is a motivated distinction between simple rules of logical entailment and rules of rational inference, at least partly why one should not always go forward in inference with trivial logical entailment. Later we'll see that higher-order awareness is the other major component of this explanation.[6]

Our second vignette is

> *The Polling Case*
> Suppose you ask 100 people on a particular street if they are Democrat or Republican, and 87 reply that they are Republican. You rationally believe that most people in the neighbourhood are Republican. Then you discover that you are standing on Coin-Flip Street: questions there are answered by coin flip. Should you retract your view that most in the neighbourhood are Republican?

A natural thought is that it depends on how you got into your initial epistemic position. If you based your belief that most in the neighbourhood are Republican on the polling data, then you should give up the view that most in the neighbourhood are Republican, for the data have been undermined by new information. If you came to believe that most in the neighbourhood are Republican on the basis of an alternative source, however – say a well-trusted political demographer told you as much – then you should hang on to your belief that most in the neighbourhood are Republican, and simply add to it

5 See Neta (2019).

6 Robert Nozick first made the point about entailment and inference in his Ph.D. thesis (Nozick, 1963). Gilbert Harman built on the point in (Harman, 1970) and (Harman, 1986).

new thoughts about Coin-Flip Street. Rationality once again looks to depend on how mental states are based on one another across time. This is why there is a motivated distinction between countervailing and undermining considerations. Put another way: this is why there is a motivated distinction between negation and indicating defeat. Later we'll see that higher-order awareness is central to the latter sort of defeat.

Our final vignette is

> *The Visual Case*
> You are in the Oval Office and it looks as if something red is on the Resolute Desk. You believe there is something red on the desk. Then you learn that lighting in the Oval is tricky: non-red things look red under the lights. Should you retract belief that there is something red on the Resolute Desk?

Here too it depends on how you got into your initial epistemic situation. If you based your belief that there is something red on the Resolute Desk on your visual impression of the desk, then you should retract your view, since visual reason has been undermined by information about tricky lighting. If you believe that there is something red on the Resolute Desk for a different reason, however – perhaps you placed a red gift there for the President – then you should not retract your belief that there is something red on the Resolute Desk. You should retain that belief and add to it new thoughts about tricky lighting. Once more rationality seems to depend on how mental states are based on one another across time. Here too we have a motivated distinction between countervailing and undermining considerations – only this time it is a perceptual state rather than a belief state which is shown not to indicate the truth.[7] We'll now see that higher-order awareness is central to indicating defeat as well.

4 Higher-order Awareness and Rationality

Something important is missing in the stories of rationality before us. In essence they presuppose a crucial ingredient in that rationality, which ingredient should be made explicit. It is true that rationality in the stories depends on how mental states are based on one another across time, but it is also true that rationality in them depends on something more than that.

7 For comparative discussion of doxastic and perceptual undercutting see (McGrath, 2021).

The missing ingredient is capacity for first-person awareness of thought. To a rough first approximation, it is capacity for first-person awareness of how mental states have been based on one another across time, the capacity for what we'll call a "*de se* running tab" of the unfolding of cognition in the recent past. For ordinary people like us, this sort of capacity is fully exercised in situations like those in the vignettes before us. Quotidian cases like them are fully soaked in its exercise, which means both the presence and functioning of a *de se* running tab is simply taken for granted.

For example: if you come to believe that there are apples or oranges in the fridge on the basis of your view that there are apples in the fridge, and you are anything like a normal person, you will have higher-order awareness that this is so. Put rather formally: you will appreciate from the first-person perspective that you have come to believe the relevant disjunctive content on the basis of belief in something which is one of that content's disjuncts. Put less formally: you will appreciate from a *de se* perspective that you have come to the apples-or-oranges belief on the basis of an apples belief. This appreciation is exercise of the relevant higher-order capacity. It involves first-person understanding of how mental states are based on one another over time. Similarly, if you come to believe that there are apples or oranges in the fridge on the basis of testimony from another, and you are anything like a normal person, you will have higher-order awareness of how you came to believe the disjunctive content. You will appreciate from the first-person perspective which of your mental states are based on which, and this too will amount to exercise of the relevant higher-order capacity.

Likewise for the Polling Case: if you came to believe that most in the neighbourhood are Republican on the basis of the polling data, and you are a normal person, you will exercise capacity for higher-order awareness of how you came to believe this about the neighbourhood. You will appreciate from the first-person point of view that you came to do so on the basis of the polling data. On the other hand: if you came to believe that most in the neighbourhood are Republican because a trusted political demographer told you as much, and you are a normal person, you will exercise capacity for higher-order awareness of this aetiology instead. You will appreciate from the first-person point of view that you came to believe as you do via the political demographer.

Likewise for the Visual Case: if you came to believe that there is something red on the Resolute Desk because it visually appears as if there is something red on the desk, and you are a normal person, you will exercise capacity for higher-order awareness of how you came to have this belief. You will know in a *de se* way that you came to do so on the basis of visual experience. But

if you came to believe that something was red on the Resolute Desk on the basis of having placed something red there for the President, and you are a normal person, you will exercise capacity for higher-order awareness of this aetiology instead. You will know in a *de se* way that you came to believe there is something red on the desk on the basis of having placed something red there yourself.

Most will agree that the relevant higher-order capacity is exercised in cases like those before us – at least when they are populated with creatures like us, creatures with ready-to-hand first-person awareness. What is not widely appreciated is that this sort of higher-order capacity is central to the *rationality* of ordinary cases like those before us. Since that is the major hypothesis of this article, our next task is to show how and why it is so. We'll see that higher-order awareness is central to three things: epistemic conservatism, the preservation of rationality across time, and the indicator defeat of rationality by new information. The first two topics are covered in the next section, the third is left to the final section of the article.

5 Epistemic Conservatism and the Preservation of Rationality

Environmental and specific reasons work differently in rational thought. Specific reasons can make it rational to form a belief and also rational to persist in believing, but environmental reasons can only play the latter role. They can only make it rational to persist in believing once belief has been sparked-off.

Environmental reasons cannot rationalize coming to believe. They are not specific enough in their content to play that epistemic role. Your environmental reasons to believe Big-Tex, for example, do make it rational to continue believing Big-Tex after you have lost touch with the specific reason you had to believe it in the first place; but your environmental reasons to believe Big-Tex do not make it rational to form a new belief that Texas is bigger than Germany. At most they make for a specific reason to think that you are a conditional indicator of Big-Tex's truth-value: if you believe it, then, in light of your environmental reasons, your belief in Big-Tex is itself reason to believe Big-Tex is true; and if you reject Big-Tex, then, in light of your environmental reasons, your rejection of Big-Tex is itself reason to believe Big-Tex is false.

In general for any claim P: specific reason to believe P is something which can make it rational to form a belief in P, and also something which can make it rational to retain belief in P; but environmental reason to believe P is something which can only make it rational to retain belief in P. This asymmetry

helps to explain three things: the ring of truth in epistemic conservatism, the preservation of rationality in the absence of specific reasons, and the undoing of rationality by indicator defeat. We tackle the first of these topics here (and leave the second to the next section).

Suppose you believe P for a specific reason. Whatever that reason turns out to be it is something which indicates the truth of P. Hence the fact that you believe P for a specific reason is *itself* a specific reason to believe that your belief in P indicates the truth of P. If your specific reason for P indicates the truth of P, after all – which it does by stipulation – and your belief that P is based on a specific reason for P – which it is by stipulation – then, in those circumstances, your belief in P likewise indicates the truth of P. Therefore, believing P on the basis of a specific reason for P is itself a specific reason for an environmental reason to believe P, namely, specific reason for the claim that one's belief in P indicates the truth of P.

This is why there is a ring of truth in epistemic conservatism. The relevant point is not that belief in P is itself a specific reason to think P true, nor that belief in P indicates the truth of P. The relevant point is that belief in P sparked-off by a specific reason for P itself indicates that P is true, and this is so even when the original specific reason for P is long forgotten. Having said that: since believing P on the basis of specific reason for P is sufficient for one's belief in P to indicate the truth of P, one might suppose that whenever one believes P on the basis of a specific reason to believe P, one has *both* a specific reason for P and an environmental reason for P. But that isn't quite right, and the point here turns out to be important.

Whenever you believe P for a specific reason, your belief in P is thereby an environmental reason to believe P. That much we've seen. It does not follow, though, that whenever you believe P for a specific reason, you thereby *possess* an environmental reason to believe P. There is a big difference, after all, between the existence of a reason to believe P and the possession of a reason to believe P. If there exists a reason to believe P but you are entirely unaware of that reason, or entirely unaware of it *as* a reason to believe P, then you do not possess a reason to believe P. In either case there may well *be* reason to believe P even though you fail to possess it. In order to possess a reason to believe P you must be aware of it *as* a reason to believe P, if only implicitly.

This means to have an environmental reason to believe P, which consists in the fact that your belief in P is based on a specific reason for P, you must be aware that you believe P for a specific reason. If you believe P for such a reason but have no clue that you do so, you fail to possess the environmental reason for P which is at your cognitive fingertips (so to say), for you fail to have any grip on the fact that you believe P on the basis of a specific reason for P. In such

a case – which we'll explore in the next section – there exists an environmental reason to believe P, and that reason consists in your cognitive situation, but the relevant facts are distal to your mind. Whenever you believe P on the basis of a specific reason for P, therefore, you will thereby have both sorts of reasons for P only if you have some kind of *grip* on the fact that you believe P on the basis of a specific reason for P.

First-person higher-order awareness of cognition is a central ingredient in the persistence of rational belief. This sort of awareness is crucial to rational belief which is not based on specific reason to think its content true. Rational belief of that sort can only occur when based on an environmental reason to believe. For belief to be so based, however, the believer must possess the environmental reason in play. A belief is rational for an environmental reason, therefore, only if the believer has some sort of grip on the fact that the belief was initially sparked-off by a specific reason to think its content is true. Once belief is created for a specific reason, it can rationally persist despite that reason withering away. Once belief is created for a specific reason, it can rationally persist even after that reason fades to obscurity. This is the thread of truth in epistemic conservatism. The phenomenon relies on higher-order awareness of thought.

In particular, an agent must be aware not only that they believe what they believe but also that they do so, at least initially, on the basis of a specific reason to believe what they believe. This is one way in which rationality like ours depends on higher-order awareness. Such awareness is central to its preservation even when our grip on specific reasons which initially generate it fade from view. It turns out that the undoing of rationality often requires higher-order awareness as well. That is our next topic.

6 Indicator Defeat

In Section 4 we noted that higher-order awareness is present in everyday belief-forming scenarios. In Section 5 we noted that higher-order awareness helps to explain how rationality persists in the absence of specific reasons. Here we forge a link between higher-order awareness and the undoing of rationality like ours.

One clear way to do this is by considering a case devised by Dorothy Edgington. Close examination of it reveals important work done by higher-order awareness in everyday rationality. Here is Edgington's case (Edgington, unpublished):

> *The Bayesian Burglar*
> You are a burglar who wants to break into a particular house. You also want not to get caught when doing so. You are confident that you will get caught if the house has an alarm, and confident that you will not get caught if the house has no alarm. You start with no clue about whether the house has an alarm, but you plan to case the joint to find out. You are sure if the house has an alarm you will see it, and if the house has no alarm you will see that too. You know the only way to detect an alarm is by casing the joint, so your plan is to decide whether to break in after doing so.

Let Alarm be the claim that there is an alarm on the house, Break-In be the claim that you break in to the house, and Caught be the claim that you are caught breaking in to the house. It is a consequence of the set-up that you start with low confidence that you break in to the house given it has an alarm, and high confidence that you break in to the house given it has no alarm. This means it is a consequence of the set-up that your initial take on whether you break in to the house sees that issue as probabilistically dependent on whether there is an alarm on the house. This seems right: since you plan to break in to the house exactly if it has no alarm, you plan to determine if the house has an alarm by casing the joint, and you're sure that you'll discover if the house has an alarm when doing so, it turns out your break-in behaviour begins tethered by your lights to whether there is an alarm on the house.

Next you case the joint and it looks to have no alarm. The look-state you go into when casing the joint perturbs your credence function, prompting an increase in confidence for no-Alarm. In turn this prompts a decision to break in. At precisely *that* point, however, something interesting should happen. Your worldview should go from one on which

$$\textbf{confidence}(\text{Break-In given Alarm}) \neq \textbf{confidence}(\text{Break-In given no-Alarm})$$

to one on which

$$\textbf{confidence}(\text{Break-In given Alarm}) = \textbf{confidence}(\text{Break-In given no-Alarm}).$$

In other words: after casing the joint and adjusting your take on things, whether or not you break in should go by your lights from being probabilistically dependent on whether there is an alarm on the house to probabilistically independent of that issue. Given your desires and goals in the case, that is a significant shift of opinion.

But why does it happen?

Lewis puts the relevant point this way (Lewis, unpublished):

> [There is] a good reason why initial confidence for Break-In given Alarm and initial confidence for Break-In given no-Alarm should differ: before you looked, you thought that if there was an alarm you would most likely spot it and be deterred, whereas if there wasn't you would probably be undeterred and go ahead. But after you've already done your looking, and revised your worldview, this reason no longer applies. There is no good reason why updated confidence for Break-In given Alarm and updated confidence for Break-In given no-Alarm should differ. Rather, they should be equal … for once you have finished looking the influence of the burglar alarm on whether you break in or not is over and done with.[8]

It is part of the set-up of the case, though, that you are a Bayesian burglar. After experience of the house introduces incoherence in your worldview – by prompting a shift in confidence for some-but-not-all propositions to which you lend confidence – you come to a new epistemic equilibrium by updating via Conditionalization or Jeffrey's rule. These are the Bayesian options, after all. Yet each of them has a well-known structural property – called *rigidity* – which ensures that whenever experience perturbs your confidence for a claim Φ and no more, then, for any claim Ψ, your new view of Ψ given Φ is identical to your old view of Ψ given Φ. For our purposes the relevant point is this: when Φ is the claim about which you change your mind on the basis of experience, updated confidence conditional on Φ remains unchanged.[9] That is always true in a Bayesian update, so it's true for a Bayesian burglar like you.

Before you looked to see if the house had an alarm, confidence for Break-in given Alarm was distinct from confidence for Break-in given no-Alarm. This was part of what made it the case that break-in behaviour was tethered by your lights to the alarm situation at the house. After you cased the joint, however, confidence for Break-in given Alarm became identical to confidence for Break-in given no-Alarm. This is part of what made it the case that break-in behaviour became untethered by your lights to the alarm situation at the house. Since updating was done with a Bayesian rule, and the Bayesian rule

8 I have brought terminology into line with this article.

9 For Conditionalization and Jeffrey's rule see the chapter on probability kinematics in (Jeffrey, 1965). For a full technical explanation of the rules, as well as discussion of why they make intuitive sense, see Sturgeon (2020), Chapters 2 and 4 respectively.

involves unchanged confidence conditional on (the conjunction of) whatever one has changed one's mind about, it follows that experience perturbed *more* than your view about whether there was an alarm on the house. Experience shifted your take on further topics as well.

Which topics?

This is where higher-order capacity enters the picture. When you see that the house has no alarm, and come to lend new confidence to Alarm on that basis, you appreciate this very fact about the unfolding of your cognition, and you do so from the first-person perspective. You have a *de se* understanding of how your own cognition unfolds. You lend confidence to no-Alarm on the basis of how things look, and likewise appreciate on the basis of doing that very thing that you have done that very thing. In creatures like us, at least, exercise of this sort of higher-order capacity permeates rationality. In the Burglar Case it explains why break-in behaviour becomes untethered by your lights to whether there is an alarm situation on the house. Higher-order awareness explains this sort of important shift of opinion.

It is difficult to say what this higher-order awareness consists in – perhaps a proprietary attitude taken to higher-order information, perhaps acquaintance with cognitive flow over time, perhaps credence lent to *de se* higher-order claims, perhaps a combination of all these things (and more). We take no stand on that here.[10] The only point needed for our discussion is that humans have some kind of recognizable capacity for a first-person grip on how cognition unfolds, some kind of capacity for understanding from the first-person perspective how mental states are based on one another in the recent past. The Burglar Case shows that this capacity plays a key role in our rational shift of opinion.

We can extend that lesson to everyday vignettes on the table. To see how consider an unusual person: Una. Suppose she is as much like us as can be save for one thing, Una lacks cognitive resources distinctive of higher-order awareness. For this reason, she lacks the capacity to have a *de se* running tab of how her recent cognition unfolds. Without seriously committing to which resources are distinctive of higher-order awareness, let's begin to spell out the scenario with the idea that Una lacks concepts needed to conceptualize her own mental life. Let's stipulate that Una's thoughts are only outwardly directed, only focused on the external world. She perceives and thinks about snow and bread and politics just as we do; and she enjoys belief about such things based

10 There is a large and heterogeneous literature on higher-order awareness. Some of it is linked to Bayesian epistemology and some of it not so much. For a good start on each sort of work respectively see Christensen (2010) and Moran (2001).

on experience of them. Una knows people say things to her from time to time just as we do; and she comes to believe things about the world based on what people say. But Una is unaware that this basing is going on. Indeed she cannot keep a running tab on how any of the basing works, for she lacks concepts needed to do so – *visual experience, belief, basing*, and so forth.[11]

Una's rationality plays out differently than ours in the everyday scenarios we've considered. Suppose it looks to her as if something red is on the Resolute Desk, for instance, and she comes to believe on that basis that something red is on the desk. Not only will Una be unaware of this basing fact in her mind, she will be incapable of conceptualizing it if brought to her attention, for she lacks the concepts needed to do so. Una is quite literally cut-off from how her own belief state is based on her visual experience. The desk visually appears to be a certain way, and she comes to believe it is that way on the basis of her visual experience, but Una has no clue any of this occurs. If you ask her about the Resolute Desk, she will insist that there is something red on it. She might even demonstrate the seen object as we would. If you tell Una about tricky lighting in the Oval Office, however, she will not understand what you are talking about, for she does not have the concepts needed to understand you.

Since Una has no conception of visual experience, she has no conception of misleading visual experience. This means Una cannot properly grasp undermining information in the Visual Case – in a recognizable sense, and despite there being a reason to retract belief in The Visual Case – Una fails to possess that reason to retract. She would possess it if she properly grasped the reason in question, if it were close enough to her intellect, so to say, but Una cannot so grasp tricky lighting by stipulation. Hence, she possesses no reason to retract belief that there is something red on the Resolute Desk. Nothing in her mind calls that belief into question.

Ordinary people should retract belief in the Visual Case. This is because news in the case triggers indicator defeat for them. But news in the case does not trigger such defeat for Una. When ordinary people receive information about tricky lighting, the news signals that they should not believe things on the basis of how they look, for how things look does not indicate how they are. Since ordinary people have a first-person grip on what they've recently based their beliefs on, news of tricky lighting joins with that first-person awareness to create epistemic pressure to retract belief based on visual experience. In this way exercise of the capacity for higher-order awareness sets the stage for

11 Nothing turns on seeing the capacity for higher-order awareness as itself pivoting on the capacity to deploy concepts. Whatever your favoured resources can be used in the arguments to follow (*mutatis mutandis*).

indicator defeat – without it no such defeat is triggered. Since Una lacks the capacity for higher-order awareness, and that sort of awareness sets the stage for indicator defeat, such defeat doesn't occur for Una in the Visual Case.

Things play out the same way for doxastic undermining. Suppose Una asks 100 people on a given street if they are Democrat or Republican, and 87 reply that they are Republican. Una duly comes to believe on this basis that most in the neighbourhood are Republican. Since Una lacks a *de se* running tab of her cognition, from her point of view salient facts in the case are that 87% of those queried said that they were Republican and that the neighbourhood is generally Republican. We might suppose that Una thinks the latter helps to explain the former, but she will be unaware that she believes either thing, much less that she believes one of them on the basis of the other. Suppose Una is then told that her polling was done on Coin-Flip Street and that questions on that street are answered by coin flip. How should she react to the news?

Una will realize that her polling data are indicative of coin flips and not political affiliation. Does this call into question, from her point of view, the Republican nature of the neighbourhood she's in? No. Does the news about Coin-Flip Street call into question, from Una's point of view, that her belief that the neighbourhood is Republican is somehow ill-based? No, for she has no concept of belief to begin with, much less a concept of ill-based belief. Since Una lacks a first-person running tab of cognition, she has no way to appreciate the link between information about Coin Flip Street and whether the relevant neighbourhood is Republican. To forge that link, *in situ*, higher-order information is needed. To accept or appreciate such information requires the capacity for higher-order awareness. Una lacks that capacity completely, so her new information about Coin Flip Street does not trigger indicator defeat. The higher-order environment needed to set-up such defeat is entirely lacking in her mind. Once Una learns about Coin-Flip Street, therefore, she should retain her view that the neighbourhood is largely Republican, reject the idea that her polling data are explained by demographics of the neighbourhood, and that is about it.[12]

From Una's position rational retraction of the view that most in the neighbourhood are Republican requires one of two things: either possession of strong evidence that it is not the case that the neighbourhood is largely Republican, or possession of strong evidence damaging to the *bona fides* of Una's reason for accepting that most in the neighbourhood are Republican. Una has no information of the first sort and cannot grasp information of the

12 Don't forget that we are talking about a variety of agential rationality. Perhaps after receipt of the news about Coin Flip Street Una's belief that her neighbourhood is Republican is itself non-agentially irrational. That idea and its denial are consistent with our discussion.

second sort. This is why she fails to possess a reason to retract her view that most in the neighbourhood are Republican. Her lack of higher-order capacity ensures that indicator defeat goes untriggered, since the stage is not set for it. Epistemic agents like us should retract belief that most in the neighbourhood are Republican, if we find ourselves in Una's position. But that is because we *find ourselves* in Una's position. Higher-order awareness is necessary to do so.[13]

Higher-order capacity is also crucial to the *bona fides* of the distinction between inference and trivial entailment. Suppose Una is told that there are apples in the fridge and comes to believe on that basis that there are apples in the fridge. Suppose Una then comes to believe that there are apples or oranges in the fridge on the basis of her recently-acquired belief that there are apples in the fridge. In the event, Una follows a trivial entailment with her belief. She is unaware this is what she is doing, of course, since she only has outwardly directed thoughts. But what happens in Una's mind is that she believes that there are apples in the fridge and then believes on that basis that there are apples or oranges in the fridge.

Suppose Una looks for an apple in the fridge because she wants one. She sees clearly that there are no apples in the fridge. Una then comes to believe on that basis that there are no apples in the fridge, replacing her old belief on the subject with its negation. From her new point of view, then, not only are there apples or oranges in the fridge, but there are no apples in the fridge. This makes it rational for Una to infer that there are oranges in the fridge after reversing her take on the apples. After all, Una has no clue about any connection between her belief in apples-or-oranges and her initial belief about apples. Higher-order capacity is needed to track that connection. This is why higher-order capacity is needed to motivate a distinction between trivial entailment and rational inference. Without such capacity it is always rational – in light of one's interests and desires, of course – to follow trivial entailment in belief, no matter where it leads. The *bona fides* of a distinction between trivial entailment and rational inference turn on higher-order capacity. Without it no such a distinction has purchase on a rational agent.

In a nutshell, then, the capacity for higher-order awareness is crucial to rationality like ours. It helps to explain how rational opinion can persist even when evidence is long gone. It guides everyday opinion and action (as in the Burglar Case). It helps to explain indicator defeat in perceptual and doxastic situations (as in Visual and Polling Cases respectively). It helps explain why trivial entailment does not always make for rational inference (as in the Fruit Case). Higher-order capacity is central to everyday rationality like ours.

13 For related discussion see Sturgeon (2014) as well as McGrath (2021).

Acknowledgments

Thanks to Tom Baker, David Christensen, Harry Golborn, Matt McGrath, Ram Neta, Matt Parrott, Maja Spener, and Alastair Wilson.

References

Christensen, David 1994. "Conservatism in Epistemology." *Nous* 28, 69–89.

Christensen, David 2010. "Higher-order Evidence." *Philosophy and Phenomenological Research* 81, 185–215.

Edgington, Dorothy (unpublished). "Tale of a Bayesian Burglar".

Friedman, Jane 2020. "The Epistemic and the Zetetic." *Philosophical Review* 129, 501–536.

Firth, Roderick 1978. "Are Epistemic Concepts Reducible to Ethical Concepts?". In: *Values and Morals*, edited by Alvin Goldman and Jaegwon Kim, Dordrecht: Kluwer Academic Publishers, 215–229.

Harman, Gilbert 1970. "Induction: A Discussion of the Relevance of the Theory of Theory of Knowledge to the Theory of Induction (with a Digression to the Effect that neither Deductive Logic nor the Probability Calculus has anything to do with Inference)." In: *Induction, Acceptance, and Rational Belief*, edited by Marshall Swain, Dordrecht: Kluwer Academic Publishers, 83–100.

Harman, Gilbert 1986. *Change in View*. Cambridge, Mass.: MIT Press.

Jeffrey, Richard 1965. *The Logic of Decision*. Chicago: Chicago University Press.

Kvanvig, Jon and Chris Menzel 1990. "The Basic Notion of Justification." *Philosophical Studies* 59, 235–261.

Lewis, David (unpublished). "Advice to a Bayesian Burglar".

McGrath, Matthew 2021. "Undercutting Defeat". In: *Reasons, Justification, and Defeat*, edited by Jessica Brown and Mona Simion, Oxford: Oxford University Press, 201–223.

Moran, Richard 2001. *Authority and Estrangement*, Princeton: Princeton University Press.

Neta, Ram 2019. "The Basing Relation." *Philosophical Review* 128, 179–217.

Nozick, Robert 1963. *The Normative Theory of Individual Choice*, Ph.D. Thesis, Princeton University.

Pollock, John 1967. "Criteria and Our Knowledge of the Material World." *Philosophical Review* 76, 28–60.

Pollock, John 1987. "Defeasible Reasoning." *Cognitive Science* 11, 481–515.

Sturgeon, Scott 2014. "Pollock on Defeasible Reasons." *Philosophical Studies* 169, 105–118.

Sturgeon, Scott 2020. *The Rational Mind*. Oxford: Oxford University Press.