Epistemic Stability

by
Nilanjan Das
B.A. (Hons.), Jadavpur University (2009)
M.A., Jadavpur University (2011)

Submitted to the Department of Linguistics and Philosophy
in
Partial Fulfilment of the Requirements for
the Degree of Doctor of Philosophy in Philosophy
at the
Massachusetts Institute of Technology
September 2016

Signature redacted

Signature of Author: ____

Department of Linguistics and Philosophy
July 6, 2016

Signature redacted

Certified By: ____

Roger White
Professor of Philosophy
Thesis Supervisor

Signature redacted

Accepted By: ____

Roger White
Professor of Philosophy
Chair of the Committee on Graduate Students

Epistemic Stability

by
Nilanjan Das

Submitted to the Department of Linguistics and Philosophy on July 6, 2016
in Partial Fulfillment of the Requirements for the Degree of Doctor of
Philosophy in Philosophy.

Abstract

I argue that knowledge and rational belief are subject to *stability conditions*. A belief that amounts to knowledge couldn't easily have been lost due to the impact of misleading evidence. A belief that is rational couldn't easily have been withdrawn upon reflection on its epistemic credentials.

In Chapter 1, I support a picture of epistemic rationality on which a belief, in order to be rational, must be *stable under reflection*, i.e., it must be capable of surviving reflective scrutiny. To make room for this condition, I defend the possibility of *higher-order defeat*, where a belief is rationally undermined by misleading *higher-order evidence*, i.e., by evidence about what one's evidence supports. I sketch an account of higher-order defeat on which higher-order evidence makes an agent's total body of evidence *fragmented*: even though a piece of evidence is available within the agent's cognitive system, the agent is unable to rationally bring it to bear upon certain questions.

In Chapter 2, I explore an analogy between knowledge and moral worth. Just as knowledge requires the agent to non-accidentally believe the truth, so too does morally worthy action require the agent to non-accidentally perform the right action. I argue that the analogy lends support to an *explanation*-based account of knowledge: a belief amounts to knowledge only if the manner in which the agent forms the belief *explains both* why the agent holds the belief (rather than losing it) and why she forms a true belief (rather than a false one). I call this view *explanationism*.

In Chapter 3, I discuss a consequence of explanationism: a belief that amounts to knowledge couldn't easily be rationally defeated by misleading evidence. This condition—*safety from defeat*—is a *stability* condition on knowledge; for it requires knowledge to involve belief that is stable under small perturbations. Safety

from defeat explains a range of different epistemic phenomena. It accounts for the explanatory role of knowledge in relation to certain kinds of behaviour, like rational perseverance. It obviates certain demanding "internalist" conditions on knowledge. It also illuminates the connection between knowledge and practical interests.

Thesis Supervisor: Roger White
Title: Professor of Philosophy

Contents

Acknowledgments

Many people have influenced the ideas presented in this dissertation. First, I owe thanks to my dissertation committee. Roger White reminded me how much small details count towards the success of philosophical projects. Julia Markovits made me see how analogies between normative concepts in ethics and epistemology could be a fruitful area of inquiry. Bob Stalnaker often understood my fledgling ideas much better than I did, and helped me articulate them clearly. Steve Yablo's unexpected challenges opened up new and exciting avenues of thought.

For much of my philosophical education, I have depended on MIT's wonderful community of epistemologists. I am especially grateful to Kevin Dorst, Sophie Horowitz, Said Saillant, Bernhard Salow, Ginger Schultheis, and Ian Wells for valuable conversations that have shaped my interests in these last five years.

For helpful comments, discussion, questions, and suggestions along the way, thanks to David James Barnett, Dylan Bianchi, Alex Byrne, Jennifer Carr, David Chalmers, Brendan de Kennessey, Nicole Dular, Kit Fine, Cosmo Grant, Lyndal Grant, Alex Gregory, Caspar Hare, Sally Haslanger, Matthias Jenny, Benjamin Kiesewetter, Jack Marley-Payne, Milo Phillips-Brown, Agustín Rayo, Miriam Schoenfield, Kieran Setiya, Jack Spencer, and Ralph Wedgwood.

This work has also been presented, in one form or other, at several places. Thanks to audiences at the MIT Work in Progress Reading Group, the MIT Dissertation Workshop, the 2015 MITing of the Minds Conference, a general colloquium at NYU, the 2016 St. Louis Conference on Reasons and Rationality, and the 2016 Humboldt Normativity Conference.

Thanks to my family—Sathi Das, Dilip Das, Rohini Chaki, and Dipanjan Das, and especially, my parents, Aditi Das and Dipak Das—for their unwavering support and their irrational optimism about the future. My gratitude to Ahona Panda defies acknowledgement: without her love and friendship, these last few years

away from home would have been intolerable.

1

Fragmented Evidence

For Hume [2000/1748], "a wise man proportions his beliefs to the evidence." Russell [2009/1948] says that "[p]erfect rationality consists, not in believing what is true, but in attaching to every proposition a degree of belief corresponding to its degree of credibility," where credibility of empirical (and in some cases, non-empirical) propositions depends on what evidence the agent possesses. Brand Blanshard [1974] explains why, from an epistemic standpoint, it makes sense to form beliefs that are well-supported by one's evidence.

> 'Surely the only possible rule', one may say, 'is to believe what is true and disbelieve what is false.' And of course that would be the rule if we were in a position to know what was true and what false. But the whole difficulty arises from the fact that we do not and often cannot. What is to guide us then?...The ideal is believe no more, but also no less, than what the evidence warrants (pp. 410-411).

These remarks lend support to a certain picture of *epistemic rationality*: from an epistemic standpoint, it is rationally permissible for an agent S to hold a doxastic attitude if and only if that doxastic attitude is well-proportioned to her evidence. Call this view *evidentialism*.[1] Evidentialism has considerable intuitive appeal: if truth is the only epistemic goal and our evidence is our best guide to the truth, then it is rational just to believe what our evidence recommends.[2]

Evidentialism just says that the limits of epistemic rationality are fixed by the agent's evidence; it doesn't say *which* parts of the agent's evidence make *which* beliefs rational. Here is a natural way of precisifying the view.

> REQUIREMENT OF TOTAL EVIDENCE. From an epistemic standpoint, an agent is rationally permitted to hold a certain doxastic attitude towards a claim P if and only if the doxastic attitude adequately reflects the degree of support P enjoys relative to the agent's *total* body of evidence.[3]

---

[1]Timothy Williamson (2000, p. 164) and Thomas Kelly (2008, §2) call this thesis a "platitude," while Earl Conee and Richard Feldman [1985] and Jonathan Adler [2002] have defended it explicitly.

[2]Three points are to be noted here. First, some writers, like Conee and Feldman [2004], understand evidentialism in a more restricted fashion than I do: for them, not only are our epistemic reasons for holding certain beliefs derived from our evidence, but our evidence cannot consist in anything but mental events. I think we need not accept this second claim in order to be an evidentialist in my sense.

Other writers, like Fantl and McGrath [2009], construe evidentialism as the view that what is rationally permissible for an agent to believe depends solely on her evidence. They question this thesis. For them, whether it is rationally permissible for an agent to believe a claim depends on whether she has *sufficient evidence* for the claim, where *sufficiency* of evidence is determined by the practical stakes in the relevant scenario. Note that the kind of pragmatic encroachment that Fantl and McGrath are after is compatible with what I am calling *evidentialism*: my version of evidentialism only says that a rationally permissible attitude has to be well-proportioned to the agent's evidence, but doesn't take a stand on what factors make an attitude well-proportioned to the evidence.

Finally, what I am calling *evidentialism* is also compatible with the position that there could be practical reasons for belief. For a recent defence of this position, see Rinard [2015]. According to the view under discussion, from an epistemic standpoint, a belief is rationally permissible if and only if it is well-proportioned to the evidence; it is consistent with this view that from a practical standpoint, it is not rational to hold beliefs that are well-proportioned to the evidence.

[3]The REQUIREMENT OF TOTAL EVIDENCE is defended by Carnap [1962] and Hempel [1965]. In epistemology and philosophy of science, it also has been assumed and defended by Salmon

In this essay, I argue that the REQUIREMENT OF TOTAL EVIDENCE is false. I begin with the observation that the REQUIREMENT OF TOTAL EVIDENCE is in tension with the possibility of *higher-order defeat*, the phenomenon whereby a belief is rationally undermined by *higher-order evidence*, i.e., evidence about what one's evidence supports.[4] In cases of higher-order defeat, the evidence on which the agent previously based her doxastic attitude is still available within her cognitive system, so her total evidence may still support the relevant doxastic attitude. Yet, the agent is no longer rationally permitted to hold the relevant doxastic attitude.

In my positive account of higher-order defeat, I argue that in such scenarios, the agent's total body of evidence is in a *fragmented* state, such that even though a piece of evidence is available to the agent, she can no longer rationally bring that evidence to bear upon certain questions.[5] The evidence, so to speak, is rendered *inadmissible* with respect to to those questions. This yields a new requirement of epistemic rationality.

THE REQUIREMENT OF ADMISSIBLE EVIDENCE. From an epistemic

[1967], Sober (1975, 2009), Adler [1989], Maher [1996], Williamson [2000], Davidson [2004], Kelly (2008, 2008) and Kotzen [2012].

One might worry that the REQUIREMENT OF TOTAL EVIDENCE turns out to be either false or without a truth-value if *epistemic permissivism* is true. For permissivists, roughly, it might be rationally permissible for an agent to hold different doxastic attitudes towards the same proposition on the basis of the same body of evidence. On some versions of permissivism, this is because there are different standards of weighing one's evidence—what Schoenfield [2013] calls 'epistemic standards'—such that relative to which there might be different doxastic attitudes that one may adopt towards the same proposition in response to the same body of evidence. So, contrary to what the REQUIREMENT OF TOTAL EVIDENCE presupposes, there might not be any unique degree of evidential support that a proposition enjoys relative to a particular body of evidence. To avoid this worry, we may simply restate REQUIREMENT OF TOTAL EVIDENCE as follows: From an epistemic standpoint, an agent is rationally permitted to hold a certain doxastic attitude towards a claim P if and only if the doxastic attitude adequately reflects the degree of support P enjoys relative to the agent's *total* body of evidence and the epistemic standard that the agent accepts.

[4]The term 'higher-order evidence' is due to Kelly [2005]. For a discussion of these issues, see, for instance, Feldman [2005], Christensen (2007, 2007, 2007, 2009, 2010), Elga [ms.], Kelly [2010], Schechter [2013], Horowitz [2014], Lasonen-Aarnio (2014, 2015) Schoenfield [2015], and Sliwa and Horowitz [2015].

[5]For a sampling of the classic sources on fragmentation, see David Lewis [1982], Donald Davidson [1982], and Robert Stalnaker [1984]. Andy Egan (2008), Agustín Rayo [2013], and Daniel Greco [2014] are amongst more recent defenders of fragmentation.

> standpoint, an agent is rationally permitted to hold a certain doxastic attitude towards a claim P if and only if the doxastic attitude adequately reflects the degree of support P enjoys relative to the body of evidence which is *admissible* with respect to the question whether P holds.

This requirement, I claim, captures the truth in evidentialism.

1.1 The Problem of Higher-Order Defeat

Consider the following examples.

> *Hypoxia.* I am a pilot flying a small aircraft to an isolated port. I have undergone a long episode of reasoning in order to determine that I have sufficient fuel to complete my journey, and found things to be satisfactory. I then receive a warning from ground control, saying that there is quite a large risk that I am suffering from a mild case of hypoxia caused by high altitude. Hypoxia makes one bad at quantitative reasoning, but leaves no detectable symptoms. I don't have any reason to distrust ground control, and I know of cases where pilots suffering from hypoxia have crashed their planes due to errors of judgement. So, I shouldn't be very confident that I have sufficient fuel.[6]
>
> *Mental Math.* My friend and I enjoy competing with each other at games of mental mathematics, attempting to solve hard math problems in our heads and then comparing our answers. My friend is as reliable as I am at such computations, and I know it. One day, we set each other a difficult problem. By reasoning reliably, I come up with the correct answer, 457. I then learn that my friend came up with

[6]This case has been discussed by many, including Christensen [2010], Elga [ms.], Schechter [2013], and Maria Lasonen-Aarnio [2014].

> a different answer, 459. I shouldn't very confident that the correct answer is 457.[7]

In each of these cases, I initially had a rational belief, which was then rationally undermined by some new evidence.[8] What's special about such instances of defeat is that the new evidence seems to have its defeating force in virtue of indicating something about what the agent's evidence supports. In *Hypoxia*, after I receive the warning, I have evidence that strongly suggests that I have made a mistake in assessing my evidence, and therefore that my evidence doesn't in fact support the conclusion I arrived at. In *Mental Math*, after I find out that my equally reliable friend came up with a different answer, I have evidence that strongly indicates that I have made a mistake in my calculation, and therefore that my evidence doesn't support the answer I came up with. Since the new evidence in each case is about the agent's own evidence, call such evidence *higher-order evidence*. Accordingly, call the kind of rational undermining brought about by such evidence *higher-order defeat*.[9]

---

[7]This example is literally borrowed from Lasonen-Aarnio [2014]. Similar cases have been discussed by Christensen [2010].

[8]Some writers, however, reject this judgement. Some of these writers belong to the faction that defends the view which is commonly known as *steadfastness* in the debate on peer disagreement in epistemology. Defenders of this view say that an agent need not revise her beliefs in the face of peer disagreement in cases similar to *Mental Math*. See Hartry Field [2000], Thomas Kelly (2005, 2010) and John Hawthorne and Amia Srinivasan [2013] for this view. For a more general form of scepticism about higher-order defeat, see Maria Lasonen-Aarnio [2014]. There are others, like Han van Wietmarschen [2013], who concede that an agent can no longer form doxastically justified beliefs in these cases, but still want to say that she remains rational in holding such beliefs.

[9]There is a non-trivial question as to which cases are instances of higher-order defeat. Consider the following example:

> *Red Wall.* You step into a room, and see that there is a red wall before you. You step out, knowing that you just saw a red wall. Your friend, waiting outside, says, "The wall was lit up with a red light that can make surfaces of any colour look red." With good reason, you believe my friend, but your friend was lying. Is it rational for you to believe, on the basis of your original experience, that the wall is red? The answer seems to be, "No."

Is this a case of higher-order defeat? According to some, it might be. For example, if we accept Williamson's [2000] *E=K* thesis, we might think that in *Red Wall*, after seeing the wall, your evidence entails the claim that the wall is red. After talking to your friend, you come to have good evidence for thinking that your evidence didn't entail the claim after all; for the new evidence strongly

The phenomenon of higher-order defeat is in tension with the REQUIREMENT OF TOTAL EVIDENCE.

Consider the first stage of *Mental Math*, when I haven't yet found out what answer my friend came up with. At this stage, I know what the math problem is, and I also know basic truths about various arithmetical operations. The conjunction of these claims entail that the right answer to math problem is 457. Next, when I learn that my friend came up with a different answer, I come to have more evidence. Even though the new evidence might affect my beliefs about what my evidence supports, it shouldn't make me lose my knowledge of what the relevant math problem is, or my knowledge of basic arithmetical truths. For example, the math problem may be written on a piece of paper right before my eyes: unless the evidence that my friend came up with a different answer miraculously makes me go blind, I should have no difficulty knowing exactly what the math problem is. Similarly, the new evidence also shouldn't make me forget how numbers are added, subtracted, multiplied, and divided: as long as I am able to perform these operations, I should be able to retain my knowledge of the basic arithmetical principles required to derive the right answer to the math problem.

It is uncontroversial that knowledge requires a very high degree of evidential support.[10] Since I originally knew the basic arithmetical truths, and what the math problem was, it is plausible that the evidential support for the conjunction of those propositions was strong. Since, even after getting the new evidence, I retain my

suggests that you didn't know the relevant claim. This, in turn, rationally undermines your belief. So, this might be thought of as a case of higher-order defeat. However, if we accept a Cartesian conception of evidence on which our evidence only consists of facts about what it's like for us, then this may be treated as a case where the defeating evidence doesn't rationally undermine the agent's evidence in virtue of suggesting anything about what the agent's earlier evidence supported. Rather, it would be treated as a case of *undercutting defeat*, where the defeating evidence just lowers the probability of the relevant claim on the agent's total evidence. See, for discussion, Feldman [2005] and Christensen [2010].

[10]The assumption that knowledge requires high evidential support is uncontroversial on most views about the relationship between knowledge and evidence. On Williamson's [2000] *E=K* thesis, any claim that an agent knows enjoys maximal evidential support—within Williamson's framework, evidential probability 1. Even others, such as Conee [2001], who don't think knowledge requires maximal evidential support, still think that known claims enjoy strong evidential support relative to the agent's total body of evidence.

knowledge, the evidential support for the conjunction of those propositions may continue to be strong. But the conjunction entails the claim that the answer to the math problem is 457. At least, under a probabilistic construal of evidential support, my evidence may support to a high degree the claim that the right answer to the math problem is 457, even after I learn that my friend disagrees with me.[11] If the REQUIREMENT OF TOTAL EVIDENCE were correct, then it could be rationally permissible for me to be highly confident in this claim. But this would make higher-order defeat impossible in this scenario; for, if the new higher-order evidence rationally undermines my belief (in accordance with the description of *Mental Math*), I cannot be very confident in this claim. This is a problem for the REQUIREMENT OF TOTAL EVIDENCE.

The problem can be generalized easily to other cases like *Hypoxia*. In *Hypoxia*, I originally knew certain facts—e.g., about my current stock of fuel, the distance to be covered, etc.—which supported the conclusion that I had sufficient fuel to complete my journey. Later, I got the evidence that I might have made a mistake in assessing what my evidence supported. It is hard to see why this evidence should affect my knowledge of the facts on the basis of which I derived my conclusion: e.g., facts about what the gas gauge of the aircraft says. What it should affect is my confidence that I accommodated all my evidence correctly. But, if I continue to know all the facts on the basis of which I derived my conclusion, then, plausibly, the evidential support for the conjunction of those propositions may indeed be very strong. Since this conjunction supports the conclusion that I have sufficient fuel, the evidential support for the conclusion may also remain strong. By the REQUIREMENT OF TOTAL EVIDENCE, I then could be rationally confident in that conclusion. This conflicts with the intuitive judgement that it is rationally impermissible for me to be confident in that conclusion any more. Thus, the problem for the REQUIREMENT OF TOTAL EVIDENCE arises once more.[12]

[11]Most evidentialists also accept this construal of evidential support. See Chapter 10 in Williamson [2000] and fn. 32 on pp. 99-100 in Conee and Feldman [2004].

[12]It is worth comparing these cases to cases like *Red Wall*, discussed in footnote 9, which may also qualify as cases of higher-order defeat. We may think that in *Red Wall*, the agent acquires conclusive evidence for the claim that the wall was red, but then the status of that information as ev-

1.2 A Trilemma

Let me now lay out the contours of the problem more clearly.

The possibility of higher-order defeat reveals a tension between the Requirement of Total Evidence and two other assumptions.

> Anti-Akrasia. It isn't rationally permissible for an agent to be confident in both a claim P *and* the claim that her evidence doesn't support P.[13]

> Evidential Opacity. It is possible for an agent's total body of evidence to support both a claim P and the claim that her total evidence doesn't support P.[14]

Both these assumptions have to be true in order for higher-order defeat to occur in examples like *Hypoxia* and *Mental Math*. In those cases, the agent has strong misleading evidence for thinking that her belief isn't well-proportioned to her total evidence. That makes it rationally impermissible for an agent to believe the claim P. Unless Evidential Opacity is true, the agent cannot have misleading evidence

idence is defeated by the new evidence about the trick lighting. So, the agent's new total body of evidence doesn't support that claim any more. In *Red Wall*, it needn't be the case that the agent's total body of evidence justifies the claim the wall evidence, while she has misleading evidence that suggests that her evidence doesn't support the relevant claim.

This is important, because this shows that the kind of account that takes care of rational defeat in *Red Wall* need not address the conflict between the possibility of higher-order defeat and the Requirement of Total Evidence that arises in *Mental Math* and *Hypoxia*. For example, we may be able to sketch a *holistic* account of evidence on which whether or not an agent possesses a piece of evidence depends on her other beliefs, i.e., her beliefs about how reliable her cognitive faculties are. So, if the agent comes to have misleading evidence that strongly suggests that her perceptual faculties are malfunctioning, then she may indeed lose the evidence she acquired by perception. We may think this is what happens in *Red Wall*. It is not clear how this would help us with cases like *Mental Math* and *Hypoxia*, where the pieces of evidence required to settle the question that the agent's belief is about remain intact. In those cases, the agent's total evidence continues to support the relevant claim, even though it seems that her belief in that claim is rationally undermined. Thanks to Robert Stalnaker for discussion here.

[13]For a general defence of this and similar principles, see Horowitz [2014]. Many writers, such as Williamson [2011], Wedgwood [2012], and Lasonen-Aarnio [2014], attack this principle.

[14]When I talk about an agent's evidence supporting P, I mean that the probability of P on the agent's evidence is greater than 0.5.

about whether her evidence supports a claim. Unless ANTI-AKRASIA is true, such evidence cannot make it rationally impermissible for the agent to hold the relevant belief.

The problem of higher-order defeat reveals that the REQUIREMENT OF TOTAL EVIDENCE, ANTI-AKRASIA, and EVIDENTIAL OPACITY cannot be true together. According to the REQUIREMENT OF TOTAL EVIDENCE, an agent can be rationally confident in a claim if and only if her total body of evidence supports it. If this is true, then ANTI-AKRASIA entails that an agent's evidence cannot support a claim P while also supporting the claim that her evidence doesn't support P, which is just the negation of EVIDENTIAL OPACITY.[15] This poses a trilemma; for it isn't obvious that any of these principles can be rejected without incurring some intuitive and theoretical costs.

1.2.1 REJECTING ANTI-AKRASIA

Consider, first, the prospects of denying ANTI-AKRASIA.[16]

Assume that in *Hypoxia*, it remains permissible for me to retain my belief that I have sufficient fuel, even after I get misleading evidence in favour of the claim that my evidence doesn't support my belief. So, ANTI-AKRASIA fails. Now, it may be very likely by my lights that my evidence doesn't support my believing that I have sufficient fuel. For example: it might be very likely that pilots with hypoxia almost always overestimate the amount of fuel they have, and I might be almost certain that I have hypoxia. If that were the case, I would have reason to be very confident that my evidence doesn't support my believing that I have sufficient fuel. Assuming that my confidence lies above the threshold for belief, I could indeed rationally believe: "I have sufficient fuel, but it is unlikely (on my evidence) that I do." There is a Moore-paradoxical quality about these conjunctions.[17] Just as it

[15]For this observation, see Alex Worsnip [forthcoming] and Maria Lasonen-Aarnio [ms.].

[16]Much of the discussion below is based on Horowitz's [2014] excellent defence of anti-akratic principles.

[17]Smithies [2012] uses this observation to argue that if an agent has propositional justification to believe P, she has propositional justification to believe that she has propositional justification to believe P. One might worry that such beliefs in fact are not incoherent: for example, a faithful

seems incoherent to assert or believe, "*p*, but I don't believe *p*," so also does it seem incoherent to assert or believe, "*p*, but it is unlikely (on my evidence) that *p*."

In response, someone might argue that knowledge is the norm of belief: one should believe a proposition *P* only if one knows *P*.[18] Since my evidence does in fact support my belief that I have sufficient fuel, I don't *know* that my evidence doesn't support the relevant belief. So, I shouldn't believe that it is unlikely (by lights of my evidence) that I have sufficient fuel. However, the response doesn't obviously succeed: it is possible to draw out a similar sort of incoherence even if we accept knowledge to be the norm of belief. Even if it isn't rationally permissible for me to hold beliefs of the form "*p*, but it is unlikely by my lights that *p*", I could still rationally be confident in such conjunctions. In *Hypoxia*, I would be willing to bet at very high odds, say at 9:1, in favour of the claim that I have sufficient fuel. But since it is also very likely that I shouldn't be so confident about the matter, I could also be willing to bet at very high odds—again, let's say at 9:1—that I shouldn't be betting on my having sufficient fuel at 9:1 odds. This behaviour seems practically incoherent.

Beyond these problems of incoherence, rejecting ANTI-AKRASIA raises a deeper problem. In *Hypoxia*, if I retain my belief that I have sufficient fuel to complete the journey, while being confident that my total evidence doesn't support this conclusion, I must also then be confident my total evidence is misleading. I might reason as follows: "Probably, I'm very lucky I'm suffering from hypoxia! Otherwise, I would have likely assessed my evidence correctly, and thus ended up believing a

---

person who in a clear-eyed manner believes in the existence of God, despite knowing that she has very little evidence for it, might indeed believe, "God exists, but it is unlikely (on my evidence) that he does." Such a belief might be irrational, but still doesn't obviously seem incoherent. In response, it is worth mentioning that even though the faithful person may indeed coherently believe the relevant conjunction, it is unclear whether the charge of incoherence couldn't arise still arise in normal scenarios, where people are committed to taking their evidence to be their primary guide to the truth. Thanks to Julia Markovits for discussion here.

[18]See, for example, Williamson (2000, pp. 255-256): "It is plausible, nevertheless, that occurrently believing *p* stands to asserting *p* as the inner stands to the outer. If so, the knowledge rule for assertion corresponds to the norm that one should believe *p* only if one knows *p*. Given that norm, it is not reasonable to believe *p* when one knows that one does not know *p*." Jonathan Sutton [2007] also argues that in order to count as justified, a belief must amount to knowledge.

falsehood!" Such reasoning seems strange.

Here is one diagnosis suggested by Horowitz [2014]: it is built into our concept of 'evidence' that our evidence is our best guide to the truth; that is precisely why it makes sense for us to proportion our beliefs to our evidence.[19] If ANTI-AKRASIA were to fail in scenarios like *Hypoxia* and *Mental Math*, then an agent couldn't always see her own evidence as a good guide to the truth. Hence, the evidentialist norm of proportioning one's beliefs to one's evidence wouldn't make any sense from the agent's own perspective. Hence, if a defender of the REQUIREMENT OF TOTAL EVIDENCE were to reject ANTI-AKRASIA, she would thereby be defending a norm of epistemic rationality which has no appeal from the perspective of at least some rational agents. Why should such rational agents then accept the REQUIREMENT OF TOTAL EVIDENCE at all? Thus, the defender of this requirement cannot reject ANTI-AKRASIA without compromising the appeal of her own position.

1.2.2 REJECTING EVIDENTIAL OPACITY

Given these arguments for ANTI-AKRASIA, we might be tempted to deny EVIDENTIAL OPACITY, i.e., the principle that even when an agent's evidence actually supports *P*, it may still support the claim that her evidence doesn't support *P*. In *Mental Math* and *Hypoxia*, the protagonist's evidence seems to have exactly this structure: even though the agent's total evidence supports a claim, misleading higher-order evidence makes her doubt whether it in fact supports that claim.

However, one may insist that this isn't really true. One may insist that all of us in fact possess *a priori*, indefeasible evidence about the evidential support relations that hold between possible bodies of evidence and different hypotheses. Given that we suffer from certain computational limitations, we are unable to determine what evidential support relations hold between which pieces of information. This may be especially plausible in cases like *Mental Math* where the relevant evidential support relation is a logical consequence relation. In virtue of knowing what various logical constants mean, we already have conclusive *a priori* evidence about all

[19] Something like this idea underwrites Blanshard's remarks quoted at the beginning of this essay.

logical entailments, but our computational limitations prevent us from exploiting such evidence in the course of deductive reasoning. So, it is just a mistake to say that we lack access to what our evidence supports in *Mental Math.*[20]

This line of reasoning doesn't seem plausible in general. Even though an agent might have *a priori* defeasible evidence for various logical truths, it is not obvious that she also has evidence for claims about non-deductive support relations that might hold between possible bodies of evidence and various hypotheses. For example, plausibly, whether or not a hypothesis is well-supported by a body of evidence depends not just on various explanatory relations between that evidence and the hypothesis, but also on the prior probabilities that are assigned to the hypothesis and the evidence. Since an agent might be rationally uncertain about what the right prior probabilities are, it is possible for her to be uncertain about the evidential support relation between a body of evidence and a hypothesis.

Moreover, even if we grant that a rational agent always has access to facts about evidential support relations, we still wouldn't be able to decisively refute EVIDENTIAL OPACITY. This is because EVIDENTIAL OPACITY postulates the possibility that an agent may have misleading evidence for the claim that her evidence doesn't support a claim. Even if an agent has perfect information about what any particular body of evidence supports, she may still have misleading evidence about what her own body of evidence *includes*. If that happens, then EVIDENTIAL OPACITY could still be true, and the problem for the REQUIREMENT OF TOTAL EVIDENCE would persist.

[20]An additional consideration might be this. Our best formal theories of epistemic rationality require rational agents to be *logically omniscient*, i.e., to be certain about all logical truths at all stages of inquiry independently of all empirical investigation. Take, for example, any probabilistic model of rational degrees of belief. Under such a model, an agent, who obeys all the rational constraints on degrees of belief, must also assign credence 1 to all logical truths. Such rational agents, therefore, must be rationally certain, independently of all empirical investigation, about what body of evidence entails what propositions. If we take these theories seriously, we may want to agree that a rational agent has *a priori* evidence about what body of evidence entails what propositions. If a rational agent has *a priori* evidence about entailment relations between propositions, why not also agree that she also has *a priori* evidence about all other kinds of evidential support relations between propositions? See, for discussion of this point, Smithies [2015], Titelbaum [2015] and Littlejohn [forthcoming].

To see this, consider a simple reliabilist conception of *evidence*, on which an agent's evidence consists of all and only information derived from reliable exercises of her cognitive faculties. Next, imagine an agent, Samantha the reliable clairvoyant, who comes to gain information that the President is in New York using her powers of clairvoyance. However, Samantha also possesses other misleading evidence suggesting that the deliverances of her clairvoyant visions are unreliable. On the simple reliabilist account of evidence, Samantha has conclusive evidence that the President is in New York, but has misleading evidence in favour of the claim that her evidence doesn't include that claim, and, given the presence of other misleading evidence, also doesn't support it. If such an account of evidence were true, this would be a scenario which makes EVIDENTIAL OPACITY true.[21] But this is in conflict with ANTI-AKRASIA and the REQUIREMENT OF TOTAL EVIDENCE.

For someone who wants to deny EVIDENTIAL OPACITY *tout court*, a natural strategy would be to accept a Cartesian picture of evidence on which an agent's evidence consists only of facts concerning her current phenomenal states, i.e., facts about *what it's like* for her at that time. It is commonly thought that such states and the absence thereof are *luminous* to the agent: if they obtain, the agent knows by introspection that they do, and when they are absent, the agent knows by introspection that they are absent. So, it becomes relatively easy to determine whether

[21]This example is adapted from Bonjour [1985]. Examples of a similar structure can be constructed under Williamson's *E=K* thesis, which says that an agent's evidence consists all and only of propositions she knows, is a good example of this. Now, there is good reason to think that both negative and positive introspection principles fail for knowledge: an agent who knows may not know that she knows, while an agent who doesn't know may not know that she doesn't know. See Williamson [2000] for an argument against the positive introspection principle for knowledge, more popularly known as the KK principle, i.e., the principle that if an agent knows, then she knows that she knows. The negative introspection principle for knowledge, i.e., the principle that if an agent doesn't know, then she knows that she doesn't know, seems more obviously false; for an agent who has a justified false belief that she knows a certain claim wouldn't know that she doesn't know, even when she doesn't know. If these introspection principles are false, then Williamson [2011] shows that it is possible for us construct examples of 'improbable knowing', where an agent knows a proposition *P*, but it is improbable on her evidence that she knows *P*. In such cases, it could also be the case that an agent's total evidence supports a certain claim *Q*, but it is improbable on her total evidence that her total body of evidence supports that claim. For arguments against Williamson's anti-KK argument, see Greco [2014], Stalnaker [2015], and Das and Salow [forthcoming].

one's evidence includes a certain proposition; all the agent has to do is introspect and see whether she in a certain phenomenal state. In this sense, the agent will have access to what her evidence does or doesn't include.

But here is the problem. Williamson's [2000] anti-luminosity argument purports to show that there is no non-trivial condition that is luminous to us in the manner suggested by the Cartesian picture. Even if we agree that facts about an agent's phenomenal states exhaust her evidence, we could construct a series of phenomenal states $S_1, S_2, \ldots, S_n$, where the agent gradually goes from a state of feeling cold to a state of not feeling cold. So, S_1 is a state where the agent is feeling cold, while S_n is a state where she is not. But, for any two consecutive states S_i and S_{i+1}, the agent cannot distinguish her feeling of coldness at S_{i+1} from her feeling of coldness at S_i, and vice-versa. However, if the agent is feeling cold at S_1 and not feeling cold at S_n, then there has to be a first state S_j where the agent is not feeling cold, but at S_{j-1} she is. But then, at S_{j-1}, the agent couldn't possibly know that she is feeling cold. *Ex hypothesi,* she cannot distinguish her feeling of coldness at S_j from her feeling of coldness at S_{j-1}. If she were to believe at S_{j-1} that she feels cold, then she only unreliably avoids error in that case; for she could easily have formed a false belief about whether she feels cold in S_j. Hence she doesn't know at S_{j-1} that she feels cold. Thus, if we like this argument, we cannot accept the Cartesian thesis about the luminosity of phenomenal states.

If our evidence isn't luminous to us in the relevant sense, then it is hard to avoid the possibility of EVIDENTIAL OPACITY. Consider an agent who holds a justified belief that she is feeling cold on the basis of her sensations of feeling cold. Now, if such an agent were to get misleading evidence that suggests that her ability to discriminate her phenomenal states are malfunctioning, then her justified belief could be rationally undermined. For example, in Williamson's example, at S_{j-1}, an agent who doesn't know that she is feeling cold could get evidence strongly suggesting that she is attributing sensations of coldness to herself by wishful thinking.[22] In

[22]It is questionable whether an agent who knows that she is feeling cold could have her knowledge defeated by such misleading evidence. For scepticism along these lines, see Hawthorne [2007].

such a scenario, even though the agent's evidence, consisting of her sensations of coldness, may strongly support her belief, she will have misleading evidence for thinking that her evidence doesn't support the content of her belief. As a result, she won't be able to rationally believe it. Thus, the problem for the REQUIREMENT OF TOTAL EVIDENCE can be resurrected even on the Cartesian picture.

1.2.3 REJECTING THE REQUIREMENT OF TOTAL EVIDENCE

If the theoretical costs of rejecting EVIDENTIAL OPACITY seem too great, our only remaining option is to give up the REQUIREMENT OF TOTAL EVIDENCE. But this option doesn't seem viable.

The REQUIREMENT OF TOTAL EVIDENCE is motivated by certain consequentialist considerations. I. J. Good [1967] claimed to have proved the following result:

> GOOD'S THEOREM. Whenever evidence is available (for gathering and use) at a negligible cost, gathering the evidence and using it in forming one's beliefs maximizes expected utility.

Suppose we assume a form of *epistemic consequentialism*, on which a certain policy of belief-formation is epistemically rational for an agent to adopt just in case that policy maximizes *expected accuracy* of her beliefs.[23] In conjunction with GOOD'S THEOREM, we get that whenever evidence is available (for gathering and use) at a negligible cost, gathering the evidence and using it in forming one's beliefs is epistemically rational; for such a policy, according to GOOD'S THEOREM, maximizes expected accuracy.[24] Why does this support the REQUIREMENT OF TOTAL EVIDENCE? Good [1967] says:

> The observations already made can be regarded as constituting a record. The process of consulting this record is itself a special kind of observation. We have justified the decision to make this observation and

[23] For discussions of this form of epistemic consequentialism, see Joyce [1998, 2009], Greaves and Wallace [2006], and Leitgeb and Pettigrew [2010, 2010].

[24] For more discussion of Good's result, see Skyrms [1990], Kadane, Seidenfeld, and Schervish [2008], and Buchak [2010].

> to use it, provided that the cost is negligible. In other words we have justified the use of all the observations that have been made, and this is the principle of total evidence. (p. 320)

So, if GOOD'S THEOREM is correct, forming one's beliefs on the basis of all the evidence one possesses maximizes expected accuracy. In fact, it is the only policy of belief-formation that maximizes expected accuracy when the total body of evidence recommends a different set of beliefs from that recommended by a proper subset of it. So, unless we can find some fault with GOOD'S THEOREM, there is good reason to accept the REQUIREMENT OF TOTAL EVIDENCE.

Let us take stock. In this section, I have shown that the problem of higher-order defeat poses a trilemma: we must reject either one of the assumptions that underwrite the possibility of higher-order defeat—ANTI-AKRASIA and EVIDENTIAL OPACITY—or the REQUIREMENT OF TOTAL EVIDENCE. It isn't obvious whether any of these principles can be rejected without certain intuitive and theoretical costs.

Some might think these costs themselves don't give us decisive reason for accepting these principles. Maria Lasonen-Aarnio [ms.] has argued that we ought to reject the intuitions underlying ANTI-AKRASIA. Alternatively, writers who are fond of iteration principles in epistemology, such as Greco [2014] and Stalnaker [2015], may also reject Williamson's anti-luminosity arguments, thereby clearing room for the denial of EVIDENTIAL OPACITY.

Others might offer pessimistic solutions. Christensen [2010] has argued that cases where the REQUIREMENT OF TOTAL EVIDENCE conflicts with ANTI-AKRASIA due to misleading higher-order evidence are in fact epistemic dilemmas, where the same agent is subject to conflicting demands of rationality. Worsnip [forthcoming], by contrast, has sketched an alternative picture on which there are two fundamentally different notions of rationality: one of these is coherence-based and validates ANTI-AKRASIA, while the other is evidence-based and validates the REQUIREMENT OF TOTAL EVIDENCE.

In what follows, I want to suggest a different solution to the trilemma posed above: we must reject the REQUIREMENT OF TOTAL EVIDENCE.

1.3 The Requirement of Admissible Evidence

Consider Good's argument for the REQUIREMENT OF TOTAL EVIDENCE. Two crucial assumptions of the argument are the following.

> PERFECT ACCESS. If the agent's total body of evidence is E, E entails that her total body of evidence is E.
>
> EVIDENT RATIONALITY. Any subset of the agent's total body of evidence entails that the agent is a perfectly rational Bayesian agent, i.e., she always performs acts that maximize expected utility relative to the evidence she uses.[25]

If either of these assumptions fails, then gathering more evidence and using it to form one's belief need not maximize expected accuracy even when that evidence is cost-free. Now, if we are persuaded by Williamson's anti-luminosity arguments *and* countenance the possibility of misleading evidence about our own rational capacity to respond to our evidence in cases like *Mental Math* and *Hypoxia*, we should indeed reject both these assumptions. This, in turn, means that proportioning one's beliefs to one's total body of evidence may not maximize expected accuracy of one's beliefs; so, even on the consequentialist view of epistemic rationality, REQUIREMENT OF TOTAL EVIDENCE may not be correct.

I want to suggest the following. In scenarios where an agent's total body of evidence doesn't entail what her total body of evidence includes or supports, even though the agent may indeed have evidence that supports a certain claim, that evidence may still be excluded or bracketed off from her deliberation, especially if she has reason to think that it is either not reliable in itself or cannot be reliably brought to bear on the subject-matter under deliberation. To see why this is plausible, we may look at another domain where evidence plays an important normative role:

[25] Geanakoplos [1989] notes these assumptions, and attempts to generalize Good's theorem to a class of information structures where PERFECT ACCESS isn't true. Still, in the relevant information structures, EVIDENT RATIONALITY must be true. So, if we are persuaded by the possibility of misleading evidence about one's own rationality, we should still reject Geanakoplos' generalization of Good's theorem. Thanks to Kevin Dorst for bringing this to my attention.

namely, legal practice. Legal deliberation is similar to deliberation about what to believe: both are truth-seeking enterprises. Since proportioning our beliefs to our evidence is the best way of discovering the truth, in both legal and doxastic deliberation, evidence is the primary determinant of outcome. But, in the legal domain, the unqualified dictum of proportioning the jury's verdicts to the total body of evidence isn't always accepted: evidence can be *admissible* or *inadmissible*, and inadmissible evidence should have no bearing on the deliberation of the jury.

When does a piece of evidence become inadmissible? Typically, all relevant evidence is admissible. However, the Federal Rules of Evidence say that, sometimes, even if a piece of evidence is relevant to the issue under deliberation, it could still be excluded from deliberation if its probative value is outweighed by a risk of "unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence" (Article IV Rule 403). Some of these considerations, of course, are pragmatic: for example, delay, waste of time, or unnecessary presentation of evidence would impede the efficiency of the court proceedings. Still, it is noticeable how the danger of prejudicing, confusing, or misleading the jury can make the legal official, in charge of the proceedings, exclude a piece of evidence. In such cases, there is a danger that the jury may fail to reach an accurate verdict about the subject-matter under deliberation, and thus may be unreliable. When a piece of evidence is declared inadmissible on these grounds, the court must take measures which prevent the inadmissible evidence from being suggested to the jury. Thus, considerations of unreliability can lead a piece of evidence from being excluded from deliberation in a court of law.[26]

[26]Here is an example. In *Old Chief vs. United States* 519 U.S. 172 (1997), John Lynn Old Chief, arrested after a "fracas" involving "at least one gunshot", moved for an order requiring the prosecution not to reveal the name and nature of his prior assault, which made him a prohibited possessor of a firearm. Old Chief's argument was that such evidence, though relevant, would have a prejudicial effect on the jury, and therefore would unduly tax the jury's capacity to hold the prosecution to its burden of proof beyond a reasonable doubt. Instead, Old Chief offered to stipulate, or concede to, the fact of the prior conviction without releasing its name or nature. The prosecution refused to join the stipulation, and the district court ruled in favour of the prosecution. Later, in an opinion authored by Justice David Souter, the Supreme Court ruled that the name of Old Chief's prior offense as contained in the official record was relevant to the question at hand, but the district court abused its discretion by admitting evidence that carried such risk of prejudice.

Something similar is true in the epistemic case. Typically, when an agent receives misleading evidence about whether her evidence supports what she believes, she comes to doubt whether she has reliably brought the available evidence to bear upon the subject-matter of her belief. As a result, she can no longer rationally rely in the exactly the same manner as earlier on the considerations that led to her belief; the evidence on which her belief was based now becomes provisionally *inadmissible.*

Take the *Mental Math* example. In this scenario, the agent does possess evidence for the claim that the answer to the math problem is 457, but cannot rationally be very confident in this claim due to the impact of the evidence that her friend got a different answer. So, the rational counterpart of this agent is required not to believe the claim that the answer to the math problem is 457. However, the total evidence available to the agent supports that answer. Yet, given that she has received misleading higher-order evidence, she must "set aside" or "bracket off" the considerations that led her to that answer earlier. Elga [2007] explains the thought in the following manner.

> Sometimes we may sensibly ask what a given agent believes, *bracketing* or *factoring off* or *setting aside* certain considerations. For example, suppose that your views on the trustworthiness of Jennifer Lopez derive from both tabloid reports and face-to-face interactions. In this case, we may sensibly ask what your views of Lopez are, setting aside what the tabloids say. To ask this is not to ask about your actual beliefs at some previous time. Rather, it is to ask what happens when we remove or extract tabloid-based information from your current state of belief.
>
> Likewise, in case of disagreement between you and a friend, we may ask what you believe, setting aside your detailed reasoning (and what you know of your friend's reasoning) about the disputed issue. In particular, we may ask who you think would likely be correct, setting that reasoning aside. By construction, the resulting belief state is untainted by ("prior to") your reasoning about the disputed issue. But since only the disputed reasoning has been extracted, that belief state still reflects your general information about your friend's abili-

ties. (pp. 489-490)

Similarly, in cases like *Hypoxia* where the available evidence may not entail, but may only non-deductively support the hypothesis that I have sufficient fuel, a similar phenomenon occurs. According to Christensen [2010], in such cases,

> The first-order evidence is not in question, and the explanatory connections between that evidence and the hypothesis that [I have sufficient fuel] remain incredibly strong. These connections, after all, do not depend on any claims about me, and the new information I learn about myself does not break these connections. I am still in possession of extremely powerful evidence [for the sufficiency of the fuel] - it's just that, in this particular situation, I cannot rationally give this evidence its due, because I cannot rationally trust myself to do so correctly. (p. 197)

Even though the evidence is still in place, and it supports the relevant hypothesis quite strongly, I can't rationally exploit that evidence in the light of the higher-order evidence. In this sense, the evidence is provisionally bracketed off; it becomes *inadmissible*.

In the legal context, when there is a danger that the jury might be unreliable at accommodating all the evidence, the jury isn't required take into account all the evidence while making its decision. Similarly, in the epistemic domain, when an agent might be or might have been unreliable at accommodating all the evidence, the agent—provisionally at least—need not be required to take into account all the evidence while forming her beliefs. So, the REQUIREMENT OF TOTAL EVIDENCE should be replaced with another requirement.

> THE REQUIREMENT OF ADMISSIBLE EVIDENCE. From an epistemic standpoint, an agent is rationally permitted to hold a certain doxastic attitude towards a claim P if and only if the doxastic attitude adequately reflects the degree of support P enjoys relative to the body of evidence which is *admissible* with respect to the question whether P holds.

To motivate the REQUIREMENT OF ADMISSIBLE EVIDENCE, I shall adopt a certain picture of reasoning and rationality.

I shall begin with the observation that agents like us often find themselves in a predicament where a piece of information that is available to them for one purpose isn't available for another. This suggests that the *total information state* of such an agent is *fragmented*: it is represented well as a set of different information states, each of which is indexed to some set of cognitive or practical tasks, and contains information to be used for the relevant tasks. Each of these information states is a *fragment*. The fragmentationist picture gives us a nice way of representing how reasoning works: many episodes of reasoning can be represented as involving transfer of information from one or more fragments to another. Such information transfer is subject to certain norms of epistemic rationality, according to which a piece of information that is available to one fragment sometimes cannot be rationally made available to another fragment. Such information, as I shall put it, becomes *inadmissible* relative to the cognitive or practical task that the latter fragment is indexed to. This in turn explains why higher-order defeat occurs.

This use of the fragmentationist framework has two crucial features. To see the first of these features, we need to distintinguish two uses that the fragmentationist picture could be put to. On the one hand, it could be used to predict and explain behaviour. For instance, the work of Andy Egan [2008] and of Elga and Rayo [ms.] on fragmentation is largely focused on this use of the model. On the other hand, it could be used to explain various kinds of epistemic evaluation. The use of the fragmentionist framework in explaining acquisition of deductive knowledge and irrationality, e.g., Stalnaker [1984], Rayo [2013], and Greco [2014], exemplifies this. My own project is of this latter kind.

The second important feature of my approach is this. Typically, even when the fragmentationist framework is used for explaining judgements about rationality, most writers treat fragmented states as paradigmatically irrational. For example, Davidson [1982] uses the fragmentationist idea to explain the phenomenon of *practical akrasia*, while Greco [2014] applies it to show why failures of iteration principles in epistemology are irrational. However, I am going to argue that, in in-

stances of higher-order defeat, the agent ends up in a *rationally mandated* state of fragmentation, where certain pieces of evidence are made unavailable for the purposes of forming certain beliefs, thus resulting in an apparently incoherent combination of doxastic attitudes. In such cases, therefore, being in a fragmented state can be rational.[27]

1.4 Fragmentation and Reasoning

It is uncontroversial that, often, we find ourselves and others in scenarios where a piece of information is available to us for some purposes, but not for others. Consider the following examples.

> *Imperfect Recall.* Jack has a neighbor he sees only infrequently. The neighbor's name is "Beatrice Ogden", and she lives in apartment 23-H. If asked "What is the name of the person in 23-H?" Jack is disposed to groan, scratch his head, mutter "I know this, don't tell me..." but be unable to answer. But if instead asked "How do you know Beatrice Ogden?", Jack is disposed to immediately reply, "She's the person in 23-H." [Elga and Rayo ms., p. 3]
>
> *Implicit Racism.* Many Caucasians in academia profess that all races are of equal intelligence. Juliet, let's suppose, is one such person, a Caucasian-American philosophy professor. She has, perhaps, studied the matter more than most: She has critically examined the literature on racial differences in intelligence, and she finds the case for racial equality compelling. She is prepared to argue coherently, sincerely, and vehemently for equality of intelligence and has argued the point repeatedly in the past. Her egalitarianism in this matter coheres with her overarching liberal stance, according to which the sexes too possess equal intelligence and racial and sexual discrimination are odious. And yet Juliet is systematically racist in most of her spontaneous reactions, her unguarded behavior, and her judgments about particular cases. When she gazes out on class the first day of

[27] This sort of approach isn't entirely without precedent. For example, Egan [2008] claims that fragmented agents perform better, epistemically speaking, than non-fragmented ones. By contrast, I show that epistemic rationality sometimes requires us to be fragmented.

> each term, she can't help but think that some students look brighter than others – and to her, the black students never look bright. When a black student makes an insightful comment or submits an excellent essay, she feels more surprise than she would were a white or Asian student to do so, even though her black students make insightful comments and submit excellent essays at the same rate as do the others. This bias affects her grading and the way she guides class discussion. [Schwitzgebel 2010, p. 532]

In both these cases, a piece of information is available to the agent, but she can't bring it to bear upon certain tasks. About forgetful Jack, Elga and Rayo [ms.] write, "He enjoys the sort of access that would be helpful when attempting to direct a letter addressed "To: Beatrice Ogden", but not the sort of access that would be helpful when going over to 23-H and calling a greeting through the door." In this case, we may surmise, the reason why the same information is accessible to Jack for some purposes, but not for others, has to do with the manner in which that information is stored in his memory.[28] In the case of racist Juliet, the best explanation of Juliet's behaviour is that the evidence that she has gathered by critically examining the literature on racial equality of intelligence is available to her for the purposes of arguing publicly, when she is expressly questioned on her stance regarding this issue. But that same evidence isn't available to her when it comes to assessing the work of her black students in the classroom.[29] These are what I shall call cases of *access limitations*.[30]

To represent such an agent, some writers have suggested that we adopt a 'fragmented' or 'compartmentalized' picture of the agent's total information state. According to this picture, an agent's total information state can be modelled as a set of different information states, each of which is indexed to some cognitive or practical task and carries information to be used for the relevant tasks. Each of these

[28]For discussion of how information retrieval from memory is dependent on contextual information, see Tulving and Thomson [1973].

[29]For this explanation in relation to implicit bias, see Egan [2011] and Madva [forthcoming].

[30]For philosophical discussion of access limitations, see Stalnaker [1991], Rayo [2013], Elga and Rayo [ms.], and Bianchi [ms.]. There are other examples of such access limitations. One prominent class of examples involve accessibility of information for motor tasks. See Milner and Goodale [2006] and Mussa-Ivaldi and Shadmehr [1994].

information states is a *fragment*. In *Imperfect Recall*, Jack's total information state could be represented as a set of two fragments–two different information states—one indexed to the task of answering where Beatrice Ogden lives, and the other indexed to the task of answering who lives at 23-H. While the first information states carries the information that Beatrice Ogden lives at 23-H, the second doesn't. In *Implicit Racism*, Juliet's total information state could be represented as a set of two fragments—two different information states—one indexed to the task of debating racial equality of intelligence, and the other indexed to the task of evaluating her students. While the first information state carries the information that there is no difference between races with respect to intelligence, the second doesn't.[31]

More formally, the picture could be presented as follows. Typically, in standard Hintikka- or Kripke-style relational structures for logics of knowledge and belief,

[31]Three points ought to be made here. First, when the fragmentationist posits different information states to explain the behaviour of these agents in different scenarios, she isn't engaged in an enterprise of armchair psychology. A fragment is not to be understood as what counts as a cognitive module within the agent's cognitive architecture in Fodor's [1983] sense. Rather, the model of information states that the fragmentationist offers is a useful way of representing the limitations that an agent may suffer from when it comes to accessing information for different purposes. Since access limitations do play an important role in psychological explanations of imperfect recall and implicit bias, these cases lend some support to the fragmentationist picture.

Second, there might be a version of the fragmentationist picture on which fragments are not just information states of the same agent, but rather just different agents. Davidson [1982] sketches an account of this sort in his discussion of practical akrasia. However, I am not willing to accept this account; for the picture of epistemic rationality that would go with such a picture would be very different from the standard picture of epistemic rationality. In particular, it seems to me that we are rationally required to try to make pieces of evidence available to us for certain purposes also available for other purposes. Under the fragmentationist picture, this would mean that if a piece of evidence is available to one fragment it should be made available to others. For more discussion of this requirement, see §5. However, it is not clear whether we are rationally required to try to make pieces of evidence available to us also available to other agents.

Third, I am going to be non-committal about how to individuate the tasks to which fragments are to be indexed. The only grip that we have over the something like fragmentation comes to cases of access limitations where an agent has a piece of information available for use for certain purposes, but not for others. Looking at such cases might tell how to individuate the tasks to which the fragements are to be indexed. This doesn't guarantee that we will have a clear set of criteria by which to do so. For more discussion of this issue, see Marley-Payne [ms.] and Bianchi [ms.]. According to another approach, the tasks themselves might be represented by clusters of questions that the agent is sensitive to in theoretical and practical deliberation . For a proposal along these lines, see Yalcin [2011].

we represent an agent's total information state using two accessibility relations: an epistemic accessibility relation K and a doxastic accessibility relation B. A possible world w^* is *epistemically accessible* for an agent at w, i.e., K-related to w if and only if it is compatible with what she knows at w. A possible world w^* is *doxastically accessible* for an agent at w, i.e., B-related to w, if and only if it is compatible with what she believes at w. On the fragmentationist picture, an agent's total information state is not represented by a pair of epistemic and doxastic accesssibility relations, but rather by multiple pairs of epistemic and doxastic accessibility relations, where each pair of epistemic and doxastic accessibility relations are indexed to a set of cognitive or practical tasks. So, for a certain task q, there will be a doxastic accessibility relation which captures what beliefs the agent is prepared to utilize for the purposes of performing that task, and an epistemic accessibility relation which captures what pieces of knowledge the agent is prepared to utilize for the purposes of performing that task. These two accessibility relations in turn would represent the fragment which is indexed to the task q.[32]

The important point is that for two different tasks q_1 and q_2, the epistemic accessibility relation indexed to q_1 might be distinct from the epistemic accessibility relation indexed to q_2, and the doxastic accessibility relation indexed to q_1 might be distinct from the accessibility relation indexed to q_2. For example, in *Imperfect Recall*, the fragment of Jack that is indexed to the task of answering where Beatrice Ogden lives can utilize the belief or the knowledge that she lives at 23-H: so, the worlds that are epistemically *and* doxastically accessible relative to this fragment are worlds where Beatrice Ogden lives at 23-H. By contrast, the fragment of Jack that is indexed to the task of answering who lives at 23-H can utilize neither the belief nor the knowledge that Beatrice Ogden lives at 23-H: so, there are some worlds that are epistemically and doxastically accessible relative to this fragment, where someone other than Beatrice Ogden lives at 23-H.

The fragmentationist picture gives us a nice way of capturing how reasoning works. Except in the most trivial of cases, good reasoning involves a complex and skilful task of answering a series of different questions. Not only must the agent

[32]This picture is formalized more rigorously in Appendix A.

know the answers to the questions that she asks in the course of her reasoning, but she must also know which questions to ask. Unless the agent does that well, she cannot bring to bear upon the question at hand all the bits of relevant evidence that are available to her for different purposes, but not together available for answering the question. Under the fragmentationist picture, this just means that the agent must know which fragments to bring into play at which stage of the reasoning. In this sense, good reasoning involves, on part of the agent, a skilful manipulation of her own information states. [33]

Suppose I am trying to figure out how to get from Harvard Square to the Museum of Fine Arts. I might start out by asking myself, "What is the subway stop closest to the museum?" Then, I recall seeing a map on which Heath Street is the closest subway stop. Now, I ask myself, "Given that Heath Street is the subway stop closest to the museum, what subway route does Heath Street fall under?" Once again, I remember that the Green Line stops at Heath Street. So, I must ask myself, "Given that Heath Street is the subway stop closest to the museum and it falls under the Green Line, how I can get to the nearest Green Line station from Harvard Square?" Once again, I know that the Red Line from Harvard Square stops at Park Street, where one can change over to the Green Line. With this information, I have a complete answer to the question about how to get from Harvard Square to the Museum of Fine Arts. At each stage, I strategically set myself a question, which I then settle with the information available to me. By the time I have answered all these intermediate questions, I will have accumulated all the information needed to answer the question which I wanted to settle in the first place.

The fragmentationist can represent this process quite well. For the fragmentationist, answering each question in this case could be a distinct cognitive task. There is a fragment indexed to that cognitive task, which carries the information required to answer the relevant question. In the above example, the fragment that is first activated is indexed to the question, "What is the subway stop closest to the museum?" In order to settle this question, I retrieve from memory the information that Heath Street is the subway station closest to the museum. I then make

[33]For discussion, see Rayo [2013].

that information available to this fragment. The fragment that is next activated is indexed to the question, "What subway route does Heath Street fall under?" In order to answer this new question, I once again make available to the second fragment the information that Heath Street falls under the Green Line The fragment which is then activated is indexed to the question, "How I can get to nearest Green Line station from Harvard Square?" Once again, I make available to this third fragment the information that the Red Line can take me from Harvard Square to Park Street, where I can change over to the Green Line. Once this answer is given, the fragment which is indexed to the question, "How can I get from Harvard Square to the Museum of Fine Arts?" comes to access all this information from these three fragments. The agent is now able to use all this information to resolve this question.

1.5 Two Kinds of Irrationality

Under the fragmentationist picture sketched in the last section, inquiry typically has two components. On the one hand, there is a *local* aspect of inquiry, under which the agent makes use of information available to each fragment for the purposes of forming belief. On the other hand, there is a *global* aspect of inquiry, which involves facts about how information is manipulated and transferred across different fragments. Distinguishing these two different aspects of inquiry helps the fragmentationist distinguish two different kinds of irrationality.

Consider the following cases. Suppose I know that there are 112 graduate students in the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT, while there are 45 graduate students in Science and Technology Studies (STS) and 67 in Linguistics and Philosophy (L&P). How many are there in all the departments put together? I believe there are 214. This of course is wrong. The right answer is 224. Though I knew the number of graduate students in each department, I have made a mistake while adding them up. Here, I can't be called rational; for I haven't accommodated my evidence well. But now contrast my predicament to that of Juliet in *Implicit Racism*. She does know that she has evidence which shows that different races don't differ in intelligence. This is mani-

fested by the masterful manner in which she marshals arguments in favour of racial equality of intelligence while discussing such issues with her colleagues in academia. And, yet, when she is teaching a class and evaluating her students, her behaviour clearly betrays a set of attitudes that are totally insensitive to such evidence.

These two kinds of failures of rationality are different. Now, in the case where I make a mistake in calculation, the information that is required to decide how many students there are in CSAIL, STS and L&P is available to me for the purposes of settling that question: I am prepared to utilize that information in settling the relevant question. But Juliet's failure need not be of this kind: she may well be correctly accommodating the evidence that is readily available to her in those circumstances, e.g., misleading evidence consisting largely of culturally transmitted testimony and base rate information that cast doubt on the academic abilities of black students.[34] The problem is that Juliet isn't even prepared to utilize the *other* evidence that she has about racial equality of intelligence when she evaluates her students; she is just insensitive to that evidence in such contexts. The first kind of irrationality is a *failure of belief-formation.* The second sort of irrationality is a kind of cognitive laziness, a failure even to bring certain kinds of evidence to bear upon the beliefs that guide one's behaviour in certain scenarios. It is a *failure of information management.*

The first form of irrationality pertains to the local aspect of inquiry, where the agent utilizes the information available to each fragment for the purposes of belief-formation. The second pertains to the global aspect of inquiry, which involves bringing information epistemically accessible for some purposes to bear upon other purposes. I shall call norms governing the first aspect of inquiry *norms of belief-formation,* and those governing the latter *norms of information management.* Let us focus on the global aspect of inquiry which pertains to information transfer. In Juliet's case, she is aware that she has evidence available to her about the racial equality of intelligence. And, yet, she doesn't bring that evidence to bear upon her assessments of her students. Under the fragmentationist picture, this failure of rationality could be understood as a failure to transfer information properly from

[34]Tamar Gendler [2011] makes this point quite forcefully.

one fragment to another. Thus, failures of information management are failures of information transfer.

To flesh out this picture in greater detail, let us begin by noting an obvious fact. It seems uncontroversial that human beings are capable of setting themselves questions and using available evidence to answer those questions. We try to remember answers to questions people ask us, and try to reason from known premises to conclusions hitherto unknown to us. This task of bringing to bear available information on new questions is a distinct cognitive task, distinct at least from tasks of belief-formation. If so, then the norms that apply to this task must also be different. What are these norms?

Intuitively, it seems that the regulative ideal that governs the operations of information transfer is this.

> INTEGRATION. For any fragment i, a piece of information E is to be made available relative to i if and only if E is available as *evidence* relative to some fragment j.

If INTEGRATION is satisfied, then there cannot be any piece of information which is available as *evidence* to some fragment of the agent, but is not made available relative to another fragment; neither can there be any piece of information which doesn't count as evidence available to any fragment, but is then made available relative to another fragment. In other words, if INTEGRATION is satisfied, all and only pieces of evidence available within the agent's cognitive system is made available to each fragment.

Can satisfying INTEGRATION be a requirement of rationality? On a widely accepted picture of rationality, requirements of rationality must be such that when they apply, we are in a position to know that they apply to us.[35] Now, agents like us only have access to a limited body of information about their own informa-

[35]James Pryor [2001] articulates such a view about rational requirements governing belief-formation: such requirements must be *usable* in deciding what to believe. He says: "If a belief-guiding recipe [of the form 'In circumstances C, believe p'] is to be *usable* in deciding what to believe, then the circumstances C it refers to must be circumstances such that the subject can tell whether they obtain, when he's following the recipe" (Pryor 2001, p. 116).

tion states. More often than not, we will be ignorant of whether or not a piece of information counts as evidence, or whether it has been made properly available throughout our cognitive system. If INTEGRATION were a requirement of rationality, it would often be the case that we are required by rationality to make certain pieces of evidence available to ourselves for certain purposes, but we are not in a position to know that we are subject to such requirements. So, according to the relevant picture of rationality, that might count as a reason to deny that satisfying INTEGRATION is a requirement of rationality. However, there are good reasons to be sceptical of such a view about requirements of rationality. For example, if we are sympathetic to Williamson's [2000] anti-luminosity argument, we may be inclined to say that there are no requirements of rationality which have application conditions that are luminous to us in this manner. So, requirements of rationality need not be such that whenever they apply to us, we are also in a position to know that they apply to us.[36]

Here is a slightly more plausible consideration against treating INTEGRATION as a requirement of rationality. Consider the *miners' puzzle.* Flood waters threaten to flood a mine, and ten miners are trapped either in Shaft A or Shaft B. We have enough sandbags to block one shaft, but not both. If we block one shaft, all the water will go into the other shaft, killing any miners inside it. If we block neither shaft, both shafts will fill halfway with water, and just one miner, the lowest in the shaft, will be killed. If we want to save as many lives as possible, what is practically rational for us to do? Surely, by our lights, the best thing to do is to *block the shaft where all the miners are.* But it seems that practical rationality cannot require us to pull off this action: since we are uncertain about where the miners are, so we don't know how to block the shaft where all the miners are. It is just not *something we can choose to do.*

The rough idea is this.

> OUGHT IMPLIES CHOOSABILITY. If rationality (practical or epistemic) requires an agent to φ, then φ-ing is *choosable* for the agent, i.e., there

[36]For arguments for such a picture of rationality, see Srinivasan [2015].

exists a task-specification Δ such that (i) it is circumstantially possible for the agent to do Δ while knowing that she is doing Δ, and (ii) the agent knows that doing Δ is a way of φ-ing.[37]

It is important to note that OUGHT IMPLIES CHOOSABILITY is compatible with the non-luminosity of rational requirements. Let's see how this works. In the scenario described in the miner's puzzle, the action-description "blocking the shaft with all the miners in it" satisfies clause (ii); for we might think, trivially, blocking the shaft with all the miners in it is a way of blocking the shaft with all the miners in it.[38] But it is ruled out by clause (i), because it isn't circumstantially possible for the agent to block the shaft with all the miners in it, while knowing that she is doing so under that description. Now, suppose that the miners are all in Shaft A. So, there is an action specification, namely "blocking the Shaft A", which satisfies clause (i); for it is circumstantially possible for us to knowingly block Shaft A. But it doesn't satisfy (ii): we do not know that blocking Shaft A is a way of blocking the shaft that contains all the miners. These seem like only natural options here. Given that these are ruled out, there is good reason to think that rationality doesn't require us to block the shaft with all the miners in it.

We might think that something similar is true with INTEGRATION. For example, consider a scenario where I have an imperfect memory and therefore can't retrieve a piece of information, or in a scenario where I am ask to prove a difficult theorem. In such a scenario, the task-specification "integrating all one's evidence" may satisfy (ii); for integrating all one's evidence may be treated as integrating all one's evidence. But it definitely doesn't satisfy (i), because it isn't circumstantially pos-

[37]Kolodny and MacFarlane [2010] offer a similar definition of choosability; the only difference is that they use epistemic possibility instead of circumstantial possibility in clause (i). As Cariani, Kaufman, and Kaufman [2013] note, this is too weak. Cariani, Kaufman, and Kaufman make the definition too strong, by requiring the agent also to know that it is circumstantially possible for her to knowingly do Δ. I think my definition strikes the right balance between the two. Even if this specification of choosability isn't correct, the hope is that something in the vicinity will be right.

[38]Here, it all depends on how fine-grained we want *ways* of performing an action to be. If we think of ways of φ-ing as the possible answers to the question, "How might an agent end up φ-ing?", we might say that blocking the shaft with all the miners in it isn't a way of blocking the shaft with all the miners in it; for this cannot really explain how an agent might end up blocking the shaft with all the miners in it.

sible for me to integrate my evidence, while knowing that I am doing so under that description. By constrast, performing certain simple tasks may indeed help me satisfy INTEGRATION. So, there may be a task-specification Δ such that doing Δ helps me satisfy INTEGRATION, and it is circumstantially possible for me to knowingly do Δ. However, in such a scenario, I just won't have the slightest clue as to what strategy might help me integrate the evidence I possess. Thus, I won't know that doing Δ is a way of integrating my evidence. Thus, integrating my evidence isn't choosable for me. If OUGHT IMPLIES CHOOSABILITY is right, I can't be required by rationality to satisfy INTEGRATION.

Now, in the miner's puzzle, even though I can't choose to block the shaft that contains all the miners, I can choose a course of action which best promotes the goal of saving the most number of lives *by lights of my evidence.* For example, if I know someone who might have more complete information about where the miners are, I can call that person, and find out where the miners are. Similarly, in the scenario where I am stuck with a bad memory or a difficult theorem, the best that I can do so in such scenarios to choose a course of action that best promotes the goal of INTEGRATION *by lights of my evidence.* The following, therefore, is true.

> REQUIREMENT OF INFORMATION MANAGEMENT. When the agent is addressing the task of information-transfer, epistemic rationality requires her to facilitate information transfer amongst the fragments in a manner that best promotes the ideal of INTEGRATION by lights of the evidence available to her relative to that task.

This explains the irrationality of Juliet. Juliet knows that she has access to evidence in favour of racial equality of intelligence: after all, she is able to reflect on such evidence and articulate arguments for her position on the basis of such evidence. Yet, while evaluating her students, she doesn't bring that evidence to bear upon her beliefs. Now, by her own lights, the best way of promoting the goal of INTEGRATION would involve bringing that evidence to bear upon her beliefs about the merits of her students. Since she doesn't do so, she fails to satisfy the REQUIREMENT OF INFORMATION MANAGEMENT.

1.6 Higher-Order Defeat Revisited

Let us now see how this picture of rationality can be used to handle cases of higher-order defeat.

1.6.1 Inadmissible Information

Take the case where I come to believe that the total number of students in CSAIL, STS and L&P is 214. Now, if I am told by a trustworthy friend that I have made a mistake in calculating the number of students in CSAIL, STS and L&P, but continue to believe that the total number of students is 214, I will be responsible for an additional failure of rationality pertaining to information transfer over and above my previous error in calculation.

Here, again, the additional failure of rationality can be understood as a violation of REQUIREMENT OF INFORMATION MANAGEMENT. When I am told that I have made a mistake somewhere in my calculation, the new evidence indicates that a piece of evidence was distorted during the process of information-transfer across fragments. Since this indicates that some piece of evidence that was available to me for some purposes has not been made available for certain other purposes, I gain evidence that I have failed to live up to the goal of INTEGRATION. Hence, in order to better promote that goal, I must intervene and set aside the information that was made available to the fragment which is indexed to the question about the total number of students in CSAIL, STS and L&P, and run the process of reasoning again. If I fail to intervene, I violate the REQUIREMENT OF INFORMATION MANAGEMENT.

In the scenario where I am told about my mistake in calculating the number of students in CSAIL, STS and L&P, my task is to correct the mistake. Since the mistake must have arisen in the course of information transfer across fragments, that process itself needs to be reversed. In order to do this, I must set aside the information that was made available to different fragments during the course of the reasoning, and go through the calculation once again. But, then, at least temporarily, the information that was made available to the last fragment active in the calcula-

tion will no longer be available to that fragment. So, the evidence available to the initial fragment activated in the course of reasoning cannot be brought to bear on the question that concerns the last fragment. In such a scenario, I must suspend judgement about the relevant question, or believe whatever the remainder of the evidence available to that fragment recommends.

What's remarkable about such cases is that the very goal of INTEGRATION makes room for setting aside information, when the agent receives evidence in favour of the claim that she has made a mistake in her reasoning. That makes it rationally impermissible, at least temporarily, for the agent to bring certain pieces of information to bear on certain questions. Such information which is thus "bracketed off" is, so to speak, rendered *inadmissible.*

> INADMISSIBLE INFORMATION. A piece of information is inadmissible relative to a fragment F if and only if the agent cannot avoid violating the rational requirements on information-management if the information is made available relative to the tasks pertinent to F.

In cases like *Mental Math,* evidence is rendered inadmissible in the relevant sense. In the first stage, I go through a process of reasoning on the basis of two different pieces of evidence: my evidence about basic mathematical truths and my evidence about what the math problem is. These two pieces of information are then transferred across a chain of fragments, so that relative to the terminal fragment in the chain, I come up with the belief that the answer to the problem is 457. When I receive evidence that my friend came up with a different answer, I interpret this as evidence for the claim that I haven't transferred information correctly from one fragment to another. Thus, it is quite likely on my evidence that I haven't lived up to the ideal of INTEGRATION. By the REQUIREMENT OF INFORMATION MANAGEMENT, therefore, I now must set aside the information that was made available to the last fragment which came up with the answer. Thus, though the evidence for correctly settling the problem may be available within my cognitive system, it is rendered inadmissible relative to the fragment concerned with the question as to what the answer to the math problem is. That is why, relative to that fragment,

I cannot rationally be confident that the answer to the math problem is 457.[39]

It is easy to notice that this treatment of higher-order defeat can be generalized to other cases. In *Hypoxia*, in order to arrive at my conclusion, I had to go through a series of steps, each of which, according to the fragmentationist, would require the activation of a fragment. At each of those steps, a new question was raised, and I had to proportion my beliefs about that question on the basis of the evidence made available to the fragment activated by that question. Once all the evidence had been accommodated in this manner, I came to believe that I had sufficient fuel to complete the journey. Since the process of reasoning was presumably long and arduous, I may not have known exactly which fragments were involved in this process. When I then receive a warning from ground control about the possibility of hypoxia, I am given evidence that strongly suggests that I have made a mistake somewhere in my reasoning. Given this evidence about a failure of INTEGRATION, the evidence that I have for the sufficiency of fuel is now rendered inadmissible. Since I can no longer rationally use the evidence that I used earlier to settle the question about sufficiency of fuel, I must now suspend judgement or be less confident about the sufficiency of fuel.

1.6.2 Evidentialism Revisited

According to the fragmentationist, belief-formation isn't the only cognitive task that we perform. We also have the task of manipulating information transfer from one fragment to another in the course of reasoning, remembering, and so on. That is why there is a distinct norm, captured by REQUIREMENT OF INFORMATION MANAGEMENT, which requires us to transfer information in a manner that best promotes the goal of rational integration of doxastic attitudes.

In this respect, the fragmentationist picture of rationality differs from the picture defended by the supporter of the REQUIREMENT OF TOTAL EVIDENCE. According to this requirement, the agent must proportion all doxastic attitudes to the *same total body* of evidence available to the agent. This would be possible if

[39]See Appendix B for a formal treatment of this process.

satisfying INTEGRATION were also a requirement of epistemic rationality. But, as I pointed out earlier, it seems that under circumstances where an agent doesn't know how to integrate her evidence, INTEGRATION isn't a choosable option for. If rationality can only require agents to perform choosable tasks, satisfying INTEGRATION cannot be a requirement of rationality.

By contrast, this picture preserves the analogy between the legal domain and the epistemic domain, which motivated the REQUIREMENT OF ADMISSIBLE EVIDENCE. In the legal context, we see a division of labour: on the one hand, there are the jurors who deliberate about the first-order subject-matters, e.g., whether or not the defendant is guilty of a certain kind, and, on the other hand, there is the judge who deliberates about which pieces of information the jurors ought to use while arriving at their verdict. Similarly, in the epistemic domain, though there is just one agent, we find a similar division of labour. In the course of her cognitive career, the agent must address two kinds of cognitive tasks: the cognitive task of forming beliefs about various subject-matters, and the cognitive task of making information available for the purposes of addressing the first task. The first task is analogous to the task performed by the jurors, while the second task is analogous to the task performed by the judge. When there is a risk that the jurors might incorrectly assess certain pieces of evidence, the rules of admissibility might require (or permit) the judge to exclude certain pieces of evidence from the deliberation of the jurors. Similarly, if there is a risk that certain pieces of evidence might be distorted in the course of information transfer, the requirements on information-management require the agent to render certain pieces of evidence unusable for the purposes of forming beliefs about certain subject-matters. This, in turn, lends support to the REQUIREMENT OF ADMISSIBLE EVIDENCE, the thesis according to which an agent is rationally permitted to hold a belief just in case its content is well-supported by the evidence that is admissible relative to the relevant subject-matter.

The REQUIREMENT OF ADMISSIBLE EVIDENCE preserves the core insight underlying evidentialism: namely, that it is an agent's evidence that makes a doxastic attitude rational for an agent to hold. However, an evidentialist might have a worry about this requirement. She might think that the norm captured REQUIREMENT OF

INFORMATION MANAGEMENT has a distinctly consequentialist flavour. It requires the agent to promote a certain goal, namely INTEGRATION, in the light of the information available to her. In this respect, it seems like an instance of a more general consequentialist norm, which requires rational agents to maximize expected value in decision-making. Many arguments have been given against such consequentialist norms in epistemology: the chief among them is the claim that such norms violate the core evidentialist insight that a belief that isn't well-proportioned to an agent's evidence isn't rational.[40]

Here is an example. Following Firth [1981], Berker [2013] asks us to imagine John Doe, a brilliant set-theorist on the verge of proving the Continuum Hypothesis. John Doe needs only six more months to do it, but he is suffering from a serious illness. Leading medical experts have diagnosed his illness and have informed him, correctly, that it is almost certain that he will die in two months' time. They have also told him that if he ignores all this evidence and dogmatically holds to the belief that he will recover, this will induce a partial recovery and will ensure that he will survive long enough to prove the Continuum Hypothesis. Now, if John Doe, by some kind of Pascalian mechanism, comes to believe that he will recover, will he be epistemically rational to hold that belief? It seems not. The problem for epistemic consequentialism is this. According to a consequentialist norm which requires the agent to maximize the ratio of true beliefs to false beliefs in one's cognitive system, John Doe indeed may be treated as rational. This, for the evidentialist, is unacceptable: John Doe has no good evidence for this claim.

Clearly, my account of rationality doesn't predict this. In this scenario, John Doe doesn't only have strong evidence for the claim that he will die, but he also has no misleading evidence which suggests that his evidence doesn't support this claim. So, the evidence that he possesses for the claim that he will die is indeed admissible relative to the question as to whether he will die. Now, if he fails to proportion his beliefs about that subject-matter to that evidence, he will be violating the REQUIREMENT OF ADMISSIBLE EVIDENCE. So, whatever version of conse-

[40]For such arguments, see Firth [1981], Jenkins [2007], Littlejohn [2012], Berker [2013] and Greaves [2013].

quentialism I am admitting into the picture of epistemic rationality doesn't have bad consequences that arise from standardly accepted versions of epistemic consequentialism.

1.6.3 Reconciling anti-akrasia with evidential opacity

The fragmentationist account of higher-order defeat, which I have sketched above, explains why a constraint like ANTI-AKRASIA is plausible after all. If an agent has evidence that strongly suggests that her evidence doesn't support a claim *P*, she has reason to think that any belief in *P* must be based on a mistake in accommodating all her evidence. Thus, by her own lights, if she were to believe *P* on the basis of such a mistake, she would be violating the ideal of INTEGRATION. So, according to the REQUIREMENT OF INFORMATION-MANAGEMENT, the agent would be rationally required to facilitate information transfer across fragments so as to avoid the formation or retention of such a belief. That is why an agent cannot rationally believe *P* while also having strong misleading evidence for the claim that her evidence doesn't support *P*.

Note that, on this picture, the reason why ANTI-AKRASIA holds has nothing much to do with EVIDENTIAL OPACITY. As a result, this picture allows us to make room for higher-order defeat brought about by different instances of EVIDENTIAL OPACITY. First of all, there are cases, like *Mental Math* and *Hypoxia*, where the agent knows exactly what her evidence includes, but has misleading evidence regarding the support relation between her evidence and a certain claim. We have already explained how higher-order defeat occurs in such cases. Apart from such scenarios, there are cases where higher-order defeat allegedly occurs because the agent gains misleading evidence regarding what evidence included (or still includes). Let me now say how this account handles such examples.

Consider Colin Radford's [1966] example of the unconfident examinee. The unconfident examinee gives hesitant answers to questions about English history in a quiz, but her answers are invariably correct. When asked, she reports sincerely that she doesn't know anything about English history. In fact, she did read a text-

book on English history: even though she retains the information she gathered from the textbook, she just doesn't remember reading the textbook. In response, we might say, "Of course, the examinee knows the answers to these questions! She simply doesn't know that she knows them." Suppose, following Williamson's [2000] *E=K* thesis, that an agent's evidence includes all and only facts that she knows. If we take the above description of the example for granted, then we might think of this example as a scenario where an agent's evidence entails certain propositions—i.e., certain facts about English history—but the agent has no evidence for thinking that her evidence supports these propositions. In fact, the agent might even has positive evidence for the claim that her evidence doesn't support these propositions. Hence, this is a case where EVIDENTIAL OPACITY is true.

Our account predicts that in this scenario, the unconfident examinee shouldn't believe the facts about English history that she can recall. By lights of the examinee, it is quite likely that the answers that she gives to questions about English history do not have the status of *evidence*. But, according to INTEGRATION, a piece of information should be made available to a fragment only if it is available as *evidence* to some fragment. So, by lights of the examinee, she would be violating INTEGRATION if she made that information available for the purposes of answering the questions in the history quiz. That makes the relevant piece of information inadmissible for the relevant purposes. Hence, the examinee cannot be rational to believe her own answers.

In effect, this allows us to give a unified account of higher-order defeat. Instances of higher-order defeat can be classified under two heads. In some cases, like *Mental Math* and *Hypoxia*, the agent's belief is rationally undermined by misleading evidence about what her evidence supports, even though the agent may know exactly what her evidence includes. In other cases, the justfication for the agent's belief is defeated by misleading evidence about what her evidence includes (or at least included). The virtue of the fragmentationist account that I have sketched is that it explains both kinds of cases in the same manner: it treats both kinds of cases as ones where a piece of information becomes inadmissible for the purposes of

forming a belief.[41]

1.7 The Bigger Picture

In this essay, I have tried to paint a picture of epistemic rationality, on which an agent need not always be required by rationality to proportion her beliefs to her total stock of evidence. Why does this picture of rationality matter? I think this picture sheds new light on the internalism-externalism debate in epistemology.

ANTI-AKRASIA is closely allied to a picture of epistemic rationality, on which whether or not a doxastic attitude counts as rational or justified from an epistemic standpoint depends on the agent's internal perspective on her own epistemic predicament. Call this view *internalism* about rationality and justification. ANTI-AKRASIA fleshes out this view in a certain way: it says that the rationality of a belief depends on what the agent is rational to believe regarding her own evidence. We have seen that denying the kind of internalism encapsulated in ANTI-AKRASIA has bad consequences: it makes an agent both doxastically and practically incoherent, and makes evidentialism unappealing from the first-person perspective. This puts pressure on us to accept internalism.

Yet, if we accept the REQUIREMENT OF TOTAL EVIDENCE, this form of internalism seems intolerable. For internalism now seems to require that an agent's evidence always accurately indicate what it in fact supports. In other words, EVIDENTIAL OPACITY must be false. As I pointed out earlier, this pushes us towards a Cartesian picture of evidence on which we have unfailing access to the contents of own evidence. However, Williamson's [2000] anti-luminosity argument suggests that an agent need not have perfect access to what her evidence includes, even when her evidence just contains facts about her current phenomenal states. This

[41]This account also explains our intuitions about *Red Wall*, discussed in footnote 9. In that example, the agent arguably has perceptual evidence that the wall was red. Then, she receives misleading evidence suggesting that her perceptual faculties were malfunctioning. According to my picture, even if the perceptually acquired information is available in the agent's memory, it is rendered inadmissible for the purposes of forming beliefs about the colour of the wall; for the agent has misleading evidence that suggests that the relevant information is not evidence after all and therefore that making it available to other fragments will violate INTEGRATION.

has motivated many to accept *externalism* about rationality and justification, the view that whether or not a doxastic attitude counts as rational or justified from an epistemic standpoint does not depend on the agent's internal perspective on her own epistemic predicament.

What has gone unquestioned in this debate is the REQUIREMENT OF TOTAL EVIDENCE. I have shown that rejecting the REQUIREMENT OF TOTAL EVIDENCE allows us to hold on to ANTI-AKRASIA without giving up EVIDENTIAL OPACITY. So, an agent's internal perspective on her own epistemic predicament could affect the rationality or the justificatory status of her beliefs, even though she lacks perfect access to what her evidence supports. Thus, we can accept internalism about rationality or justifcation without embracing the Cartesian picture of evidence.

Appendix A: A Formal Model of Fragmentation

Under the fragmentationist picture that I am using, a fragment is an information state which is activated in an agent when she faces certain tasks. For simplicity, I will represent each fragment using a characteristic tasks. Here is the model under question.

> *A Simple Model.* The information state of an agent is modelled as a quintuple $\langle W, T, Q, K, B\rangle$ where
>
> 1. W is the set of "worlds" or "states";
> 2. T is the set of times;
> 3. Q is a set of cognitive or practical tasks $q_1, q_2, \ldots$;
> 4. K is a partial function function from tasks, worlds, and times to sets of worlds, such that the set of worlds $K(q_i, w, t)$ represents the set of worlds that are *epistemically* accessible relative to the task q_i in the world w at time t; and
> 5. B is a partial function from tasks, worlds, and times to sets of worlds, such that the set of worlds $B(q_i, w, t)$ represents the set

> of worlds that are *doxastically* accessible relative to the task q_i in the world w at time t.

For any world w, time t, and task q_i, if $K(q_i, w, t)$ and $B(q_i, w, t)$ are defined, $K(q_i, w, t)$ is the set of worlds that are compatible with what the agent knows for the purposes of addressing the task q_i in w at t, while $B(q_i, w, t)$ is the set of worlds that are compatible with what the agent believes for the purposes of addressing the task q_i in w at t.

For any world w, time t, and task q_i, if $K(q_i, w, t)$ and $B(q_i, w, t)$ are defined, then they are subject to the following constraints:

> *Factivity.* $w \in K(q_i, w, t)$.
> *No Belief in Impossibility.* $B(q_i, w, t) \neq \emptyset$.
> *Knowledge Entails Belief.* $B(q_i, w, t) \subseteq K(q_i, w, t)$.

This model, though too simple to capture the complexity of how fragmentation really works, is a good tool for representing how a piece of information that is available for addressing certain cognitive or practical tasks may be unavailable to addressing others.

We can now characterize information transfer amongst fragments as follows.

> *Information Transfer.* When, in a world w between t_0 and t_1, a piece of information E is transferred to a fragment indexed to q_i, the information that is doxastically accessible at t_1 to the q_i-fragment will be $B(q_i, w, t_1) = B(q_i, w, t_0) \cap E$, where $B(q_i, w, t_0)$ is the set of worlds that were previously doxastically accessible to that fragment.

In the good case, where a piece of known information is transferred not just correctly, but also reliably, to the q_i-fragement, the set of worlds that are epistemically accessible to the q_i-fragment at t_1 will be $K(q_i, w, t_1) = K(q_i, w, t_0) \cap E$, where $K(q_i, w, t_0)$ was the set of worlds that were previously epistemically accessible relative to that fragment.

Appendix B: Dynamics of Higher-Order Defeat

In addition to fragments that answer various questions about the world and performs various practical tasks, I am going to assume that there is a fragment in the agent which is responsible to transferring information from one fragment to answer. And, in cases of higher-order defeat, it is this fragment which, in the light of the higher-order evidence, causes certain pieces of information to be set aside from the agent's deliberative processes. Call this fragment the *Transferrer Fragment.*

Every scenario of higher-order defeat has three stages:

> *Stage 1.* This is the stage prior to receiving the defeating evidence.
> *Stage 2.* This is the stage at which the Transferrer Fragment receives the higher-order evidence.
> *Stage 3.* This is the stage at which some piece of information is rendered inadmissible and then excluded from the deliberation of certain fragments.

In *Mental Math,* we may imagine that there are at least three fragments. The first fragment is indexed to the task of deciding what the math problem is, and what the basic arithmetical truths. The second fragment is indexed to the task of deciding what the answer to the math problem is. And the third fragment is the Transferrer Fragment indexed to the task of transferring information from one fragment to another. Call the tasks to which these three fragments are indexed to q_1, q_2 and q_3 respectively.

Let us assume that the math problem is X. For our purposes, we need three sentences:

> p is the conjunction of (i) the sentence which says that the math problem is X and (ii) the axioms of Peano arithmetic.
> a is the sentence that says that the answer to the math problem the agent is doing is 457; and
> s is the metalinguistic sentence that the sentence a is derivable from the sentence p in the language that the agent adopts;

Now, p and $\neg a$ are logically inconsistent. Hence, there are six complete truth-value assignments to the three sentences. Accordingly, the set of all possible worlds $W = \{w_1, w_2, w_3, w_4, w_5, w_6\}$ such that

> w_1 is the world where p is true, a is true, and s is true.
> w_2 is the world where p is true, a is true, and s is false.
> w_3 is the world where p is false, a is true, and s is true.
> w_4 is the world where p is false, a is false, and s is true.
> w_5 is the world where p is false, a is true, and s is false.
> w_6 is the world where p is false, a is false, and s is false.

Of these, w_1 is the actual world @.

Two assumptions are necessary.

1. *E=K.* Following Williamson's [2000] *E=K* thesis, I shall assume that every proposition that an agent knows is part of her evidence.

2. *Knowledge Loss.* I shall also assume that an agent can rationally lose knowledge over time. In other words, the set of worlds that are epistemically accessible relative to a fragment can expand over time.[42] For my purposes, I shall just assume that knowledge defeat is possible.

Let us see how higher-order defeat in *Mental Math* can be represented within this model. At stage 1, when the time is t_1, the worlds that are epistemically accessible relative to each fragment at @ are as follows:

(1) $K(q_1, @, t_1) = K(q_2, @, t_1) = K(q_3, @, t_1) = \{w_1\}$.

[42]There has been some work of non-monotonic belief-revision: for a survey, see Hansson [1999]. The non-monotonic changes in the epistemic accessibility relation in this model could also be understood in terms of contraction: for example, if the database of each fragment is contrued as a consistent set of sentences (which may or may not be closed under logical consequence), then the epistemic accessility relation of that fragment will change non-monotonically only if the database contracts. More recently, some philosophers have questioned whether knowledge can at all be defeated. See, for example, Maria Lasonen-Aarnio [2010] and Baker-Hytch and Benton [2015].

Roughly, at this stage, all the fragments have the same epistemic accessibility relation which connects w_1 to itself. So, relative to each fragment, the agent knows the same information: namely, the axioms of arithmetic, the claim that the math problem is *X*, and the claim that the sentence *a* is derivable from *p*. (This need not necessarily be the case, but it simplifies things by abstracting away from inessential messiness.) Since the first two claims entail the claim that the answer to the math problem is 457, the sentence *a* is also true in the worlds that are epistemically accessible relative to these fragments.

At stage 2, when the time is t_2, the worlds that are epistemically accessible relative to each fragment at @ are as follows:

$$(2)\ K(q_1, @, t_2) = K(q_2, @, t_2) = K(q_3, @, t_2) = \{w_1, w_2\}.$$

At this stage, again, all the fragments have the same epistemic accessibility relation which connects w_1 to itself. So, relative to each fragment, the agent has the same information as in the previous stage, except insofar as she is now uncertain about whether the sentence *a* is derivable from *p*. This is because she receives misleading higher-order evidence, which makes her uncertain about whether what she knew in fact supports the answer she came up with.

At stage 3, when the time is t_3, the worlds that are epistemically accessible relative to each fragment at @ are as follows:

$$(3)\ K(q_1, @, t_3) = K(q_3, @, t_3) = \{w_1, w_2\}, \text{ but } K(q_2, @, t_3) = W.$$

At this stage, only the fragments indexed to tasks t_1 and t_2 retain the information about what the math problem is and what the basic arithmetical truths are. But the fragment indexed to the task of getting the answer to the math question is uncertain about what the math problem was. This is because that information has now been rendered inadmissible. So, the agent cannot also be confident about what the answer to the math problem is.

2

Knowledge and Moral Worth

I wish to treat two questions together. The first is a question in epistemology: What makes a true belief knowledge? The second is a question in moral philosophy: What makes a morally right action praiseworthy? According to one proposal, both these questions admit of the same (partial) answer: an *anti-luck* condition. If an agent knows, it cannot be a matter of luck that she believes the truth. If an agent's action is morally praiseworthy, i.e., has *moral worth*, it cannot be a matter of luck that she performs the right action. This is the analogy between knowledge and moral worth. In this chapter, I explore how this analogy helps us make progress in epistemology.

In typical examples of knowledge-destroying epistemic luck, an agent forms a true belief, but the manner in which she forms her belief only unreliably leads to the truth in those circumstances. This supports the hypothesis that a belief can be free from knowledge-destroying epistemic luck only if the manner in which the

agent forms the belief—i.e., its *basis*—reliably leads her to the truth in the relevant circumstances. Is such reliability also *sufficient* for blocking knowledge-destroying epistemic luck? Some say, "Yes." This is the view we may call *reliabilism*.

> EPISTEMIC RELIABILISM. A true belief is free from knowledge-destroying epistemic luck if and only if the basis of the belief reliably leads to the formation of true beliefs in the relevant circumstances.

Using the analogy between knowledge and moral worth, I show that EPISTEMIC RELIABILISM should be rejected for the same reasons which render reliabilism about moral worth unacceptable.

Then, drawing upon the work of Nomy Arpaly and Julia Markovits, I argue that there is an anti-luck condition on moral worth which fares better than reliabilism: namely, *moral explanationism*. On this view, an action is free from luck that undermines moral worth just in case the motivating reasons underlying the action *explain both* why the agent performs a right action rather than a wrong one, and why the agent performs the relevant action rather than refraining from performing it. If we take the analogy between knowledge and moral worth seriously, the following anti-luck condition on knowledge becomes salient.

> EPISTEMIC EXPLANATIONISM. A true belief is free from knowledge-destroying epistemic luck if and only if the basis of the belief *explains both* why the agent holds a *true* belief rather than a *false* one, and why the agent *possesses* the relevant belief rather than *lacking* it.

Not only does this view avoid the problems for EPISTEMIC RELIABILISM, but it also illuminates certain structural features of knowledge. It explains why knowledge seems to require the elimination of relevant alternatives. It also vindicates the thought that knowledge requires both reliability and stability.

2.1 THE MORAL WORTH ANALOGY

Let us explore the analogy between knowledge and moral worth more carefully.

Consider the following Gettier example.

> *Fake Sheep.* Nina is looking at a field, wondering whether there is a sheep out there. At that time, an object that looks like a sheep comes into view. Thinking that it's a sheep, Nina comes to believe that there is a sheep in the field. But, in fact, the object that she sees is not a sheep, but a dog camouflaged as a sheep! However, there happens to be a sheep in a different part of the field which she cannot see.[1]

Nina seems lucky to believe the truth. That is why she lacks knowledge.[2] It is tempting therefore to impose on knowledge an *anti-luck condition*, a condition that precludes the kind of luck instantiated in examples like *Fake Sheep*.

Moral praiseworthiness, or moral worth, seems to be subject to a similar anti-luck condition. If an agent only does the right thing as a matter of luck, she doesn't deserve praise for her action; her action lacks positive *moral worth*.[3] Kant [2012/1785] offers an example that makes this manifest.

> For example, it certainly conforms with duty that a shopkeeper not overcharge an inexperienced customer, and where there is a good deal of trade a prudent merchant does not overcharge but keeps a fixed general price for everyone, so that a child can buy from him as well as everyone else. People are thus served honestly; but this is not nearly enough for us to believe that the merchant acted in this way from duty and basic principles of honesty; his advantage required it; it cannot be assumed here that he had, besides, an immediate inclination toward his customers, so as from love, as it were, to give no one preference over another in the matter of price. Thus the action was done neither from duty nor from immediate inclination but merely for purposes of self-interest. (4:397)

Where Kant speaks of *conformity with duty*, we can speak of *moral rightness*. Even though the shopkeeper's actions in this case are right, they are not morally praise-

[1] This example is borrowed from Chisholm [1989, p. 98].

[2] A diagnosis of this kind was first offered by Peter Unger [1968] in the wake of Gettier's [1963] paper.

[3] An action has positive moral worth just in case it is morally praiseworthy. Henceforth, by 'moral worth', I will mean *positive moral worth* or *moral praiseworthiness*.

worthy; for the shopkeeper has only *luckily* seems to have performed the right action in this scenario.[4] We might be inclined, therefore, to impose on moral worth an *anti-luck condition*, a condition that precludes the kind of luck instantiated in Kant's shopkeeper example.

Let us get clear about what this anti-luck condition involves. Moral rightness, arguably, can be of two kinds: *objective* or *subjective*. What makes an action objectively right for an agent to perform are facts about the consequences of the actions, or the agent's practical scenario, which the agent might not know. By contrast, the subjectively rightness of an action depends on what she has reason to believe.[5] Suppose your friend has a headache, and you have good reason for thinking that some pills you possess are painkillers. Unbeknownst to you, however, they contain cyanide. Here, it is *subjectively* right for you to give your friend those pills; for you know that your friend is in pain, and have good reason to believe that the pills you have could alleviate that pain. But it isn't *objectively* right for you to give the pills to your friend; for they will kill your friend. Now, the kind of accidentality that prevents the shopkeeper's actions from having moral worth isn't accidental *objective* rightness.

To see why, imagine that I have donated a substantial part of my salary to a charity from a motive of benevolence, but the list from which I randomly selected the charity comprised names mostly of corrupt charities. Luckily, however, the charity I selected was in fact not corrupt and utilized the donated amount in the best possible way. In this case, my action was only accidentally objectively right; for I could easily have donated my money to a corrupt charity. But that doesn't undermine the moral praiseworthiness of my action. Why? In this case, donating part of my salary to the charity is not only objectively, but also subjectively right; for, by

[4]For the discussion of this aspect of moral worth, see Herman [1981], Baron [1995, Chapter 4], Stratton-Lake [2000, Chapter 3], Arpaly [2003, Chapter 3], and Markovits [2010].

[5]For the distinction between subjective and objective rightness, see Russell [1910], Prichard [1932], Ross [1939] and Carritt [1947]. Some of these writers take subjective rightness to be sensitive to what the agent in fact believes, rather than what she has reason to believe. More recent writers, like Thomson [1986], Parfit [ms.] and Kolodny and MacFarlane [2010], take subjective rightness to depend on what the agent has reason to believe. For dissent from this account, see Smith [2010].

my lights, my donation could have helped some people in need. Given that I was motivated by this consideration, there is a non-accidental connection between my motive and the subjective rightness of my action. That is why my action has moral worth. So, for an action to have positive moral worth, it cannot be a matter of luck that the agent performs a *subjectively* right action.

This fits our verdict about Kant's shopkeeper example. In this scenario, selling his goods at a fair price is the subjectively right thing for the shopkeeper to do; for, presumably, by his lights, that price is fair. But the shopkeeper isn't motivated by this consideration, but rather by self-interest. If self-interest required him to be dishonest with his customers, he wouldn't sell his goods at a fair price. Since there is no non-accidental connection between the shopkeeper's motive and the subjective rightness of his actions, his actions lack moral worth.

To sum up, the analogy between knowledge and moral worth lies in this: for a belief to count as knowledge, it cannot be a matter of luck that the agent believes the truth; similarly, for an action to have moral worth, it cannot be a matter of luck that the agent performs the subjectively right action.

Before we proceed, let me consider an initial worry that one might have about the analogy between knowledge and moral worth. Both knowledge and moral worth involve a kind of non-accidental or non-lucky success. The success that is relevant to knowledge consists in believing the truth, while the success that moral worth requires consists in performing the subjectively right action. Truth and subjective rightness are quite different from each other: as we have seen, *subjective rightness* is an information-dependent notion, but *truth* obviously isn't. If we are looking for an epistemic notion that resembles the notion of moral worth even with respect to the kind of success it involves, then a better epistemic analogue of moral worth would be *doxastic justification*. For a belief to be doxastically justified, not only must the agent's evidence adequately support the content of the belief and therefore make the belief *subjectively right* for the agent to have, but the agent must also hold the belief on the basis of that very body of evidence. So, if an agent only holds an evidentially well-warranted belief accidentally, say, by wishful thinking, her belief wouldn't be doxastically justified. Therefore, for a belief to be doxasti-

cally justified, it cannot be a matter of luck that the agent holds the subjectively right belief. If doxastic justification is the right analogue of moral worth, why take the analogy between knowledge and moral worth seriously at all?

I have two responses. First of all, for reasons that I shall discuss in Section 4.2, I don't think doxastic justification is a good epistemic analogue of moral worth. Second, even if we grant that doxastic justification is a better epistemic analogue of moral worth, the analogy between knowledge and moral worth could still be theoretically significant. In this essay, I am concerned with finding the right anti-luck condition on knowledge; I am asking, "Given that a belief is successful insofar as it is true, what anti-luck condition must it to satisfy in order to be knowledge?" That is why it is useful for me to also ask, "Given that an action is successful insofar as it is subjectively right, what anti-luck condition must it satisfy in order to be praiseworthy?" So, my concern lies not with the kind of success that knowledge and moral worth involve, but rather with the kind of anti-luck condition required to make such success non-accidental. In this respect, the analogy between knowledge and moral worth does prove useful: it reveals what is wrong with EPISTEMIC RELIABILISM, and motivates an alternative anti-luck condition that I call EPISTEMIC EXPLANATIONISM.

2.2 RELIABILISM

Recall *Fake Sheep*. In that scenario, Nina comes to believe that there is a sheep in the field, after seeing a camouflaged sheepdog. But the sheep might easily have escaped through a hole in the fence; then, Nina would falsely believe in exactly same manner that there was a sheep in the field. So, the basis of Nina's belief doesn't reliably lead to the truth in these circumstances. That is why she seems lucky to have formed a true belief. This line of reasoning suggests that a reliability condition is necessary for blocking knowledge-destroying epistemic luck. Some may want to say something stronger: not only is reliability necessary for blocking knowledge-

destroying epistemic luck, but it is also sufficient.[6] This latter view—EPISTEMIC RELIABILISM—is our target.

The notion of *reliability* has been glossed in many ways.[7] According to most of these glosses, reliability, or truth-conduciveness, is a property of the *basis* of a belief, i.e., the manner in which the agent forms her belief. My conception of *basis* is quite liberal: the basis of a belief consists in those epistemically significant factors on which the formation of the agent's belief causally depends. Since our assessments of knowledge usually depend both on facts about the agent's evidence and the way she bases her belief on that evidence *as well as* facts about the external environment in which the agent forms her belief, the basis of a belief should include facts of both kinds.[8] For example, imagine a scenario where I walk into a room, and undergo an experience as of a red wall. If now I form a belief that there's a red wall in front of me, the basis of my belief may not only include features of the process that brought about my experience, and then led to the formation of my belief, but also features of my environment, e.g., the presence of the wall and the ambient lighting conditions.

To fix ideas, I will focus on an interpretation of reliability that I find most natural and easy to work with, namely *safety from error*. Let a *case* be a centred possible world $\langle w, s, t \rangle$, where w is a metaphysically possible world, s is an agent, and t is a time. According to the modal interpretation of *safety* that I am going to use, the basis of a belief reliably leads to the truth just in case there is no sufficiently similar case where the agent forms a belief on the same basis, but the belief is false. This gives us the following necessary condition on knowledge.[9]

[6]Such a claim has been implicitly defended by many writers in relation to cases of epistemic luck, e.g., Dretske [1971], Armstrong [1973], and Goldman [1976]. For a more explicit defence of this claim, see Pritchard [2005], who takes a reliability condition—namely, safety from error—to be the sole anti-luck condition on knowledge.

[7]For a survey of the various glosses of reliability, see Goldman [2011].

[8]For this way of thinking about the basis of a belief, see Williamson [2009].

[9]Let me say why I find the safety-theoretic interpretation of reliability quite general in comparison with two other rival conceptions of reliability: *sensitivity* and *relevant alternatives theory*.

According to Nozick's [1981] *sensitivity*-based conception of reliability, the basis of a belief is reliable in a certain scenario if and only if the agent's belief is sensitive to the truth, i.e., the agent wouldn't believe the same claim on the same basis if the claim were false. According to a diagnosis

SAFETY FROM ERROR. A belief formed on a certain basis counts as knowledge only if, in every sufficiently similar case where the agent forms a belief on the same basis, the relevant belief is true.

In *Fake Sheep*, there is a sufficiently similar case where Nina forms a belief on the same basis as in the actual case, but her belief isn't true because the sheep escaped through the fence. Due to such risk of error, Nina's belief fails to be knowledge.

If we are tempted by reliabilism in the case of knowledge, reliabilism might also appeal to us in relation to moral worth. With respect to Kant's example, we might say: the shopkeeper is lucky to perform the right action, because the manner in which he performs his action doesn't reliably lead to the right action this scenario. What motivates the shopkeeper to sell his goods at a fair price is his concern for profit. So, if the shopkeeper could have profited without selling his goods at a fair price, he would have done so. Thus, the motive of self-interest underlying the shopkeeper's actions could easily have led him astray. Therefore, we might think that a reliability condition will be necessary and sufficient for blocking the kind of moral luck that undermines moral worth in this scenario.

MORAL RELIABILISM. An action is free from worth-destroying moral luck if and only if the basis of the action reliably leads to the right action.

Let me flesh out the reliability condition on moral worth a little more carefully.

In the case of knowledge, we were concerned with the basis of the belief under evaluation, i.e., the manner in which the agent forms her belief. In the practical do-

---

defended by Sosa [1999], the safety-theoretic conception of reliability captures the same insight that underlies sensitivity: namely, that the basis of a belief is reliable in a certain scenario if and only if she wouldn't believe the claim on the same basis unless the claim were true. This counterfactual could interpreted in two ways: one reading yields sensitivity, while the other yields safety.

According to the *relevant alternatives theory* defended first by Goldman [1976], the basis of a belief is reliable in a certain scenario just in case the basis of her belief enables her to discriminate the truth of the claim she believes from relevant alternatives where the claim is false. The safety-theoretic conception of reliability captures the insight underlying the relevant alternatives theory. In the safety-theoretic framework, the relevant alternatives just are the nearby cases where the agent's belief is false. So, an agent can only reliably avoid error just in case the basis underlying her belief rules out all those cases.

main, the analogous role will be played by the *basis* of the action under evaluation, i.e., the manner in which the agent performs the action. This will include those practically significant factors on which the performance of the action causally depends. Unlike our assessments of knowledge, our assessments of moral praiseworthiness don't depend on facts about the agent's external environment, but solely on facts about her patterns of practical motivation. Accordingly, the basis of an action will consist in a practically significant subset of the *motivating reasons* for that action, i.e., the facts which causally explain the agent's action in a manner that essentially brings into play the agent's capacities of practical reasoning. Motivating reasons, in this respect, are distinct from other kinds of facts that might explain an action. For example, I may snap at you because I am tired, but my state of fatigue won't be a motivating reason for my action insofar as it doesn't explain my action by figuring in any process of practical reasoning. However, my desire that you stop talking and my belief that you will stop talking if I snap at you, which play a role in the practical reasoning that gives rise to my action, can be motivating reasons for my snapping at you.[10]

Keeping this in mind, we may state the reliability condition on morally worthy action as follows.

> SAFETY FROM WRONGNESS. An action performed on a certain basis has moral worth only if, in every sufficiently similar case where the agent acts on the same basis, the relevant action is morally right.

SAFETY FROM WRONGNESS seems right at least about Kant's example of the shopkeeper. Since the motive of self-interest could easily have led the shopkeeper to perform the wrong action, his actual action isn't safe from wrongness.

[10]Let me flag two assumptions. First of all, I am assuming here that motivating reasons are psychological states, following Davidson [1963] and Smith [1987]. But this is contentious. For opposition, see Dancy [2002]. The example is borrowed from Markovits [2014]. Second, I am also assuming that only a subset of the agent's motivating reasons will be part of the basis. Under a Kantian framework, this may well turn out to be the *non-instrumental* motivating reasons for the action, i.e., the agent's concern for ends that she values for their own sake and not as a means to some further end. For discussion, see Markovits [2010].

SAFETY FROM ERROR and SAFETY FROM WRONGNESS yield the right prediction about *Fake Sheep* and Kant's shopkeeper example respectively. However, they need not make a prediction about every case that we may imagine.

Focus on the epistemic domain. In order to determine whether or not a belief is safe from error, we would need to determine what the *basis* of the belief is. In this respect, the reliabilist proposal suffers from a problem. A belief is generated by a certain token causal process, a concrete process which takes place at a certain time and place within a certain possible world. While checking whether a belief is safe from error, we cannot simply take the token causal process as the basis of the belief: the basis of the belief has to be a causal process type which could be instantiated at different places and times across different possible worlds. However, there could be several process types corresponding to the same token causal process, some more coarse-grained than others. It is not clear how coarse-grained a process-type has to be in order for it to count as the basis of the belief; if we specify the basis of the belief too coarsely, any belief may suffer from a risk of error. This is sometimes called the *generality problem.*[11]

A similar problem can arise in the practical domain. In order to determine whether or not an action is safe from wrongness, we would need to determine what the *basis* of the action is. Suppose an agent, like Kant's shopkeeper, who is motivated solely by self-interest. She know that, in her situation, what self-interest requires of her coincides with the demands of honesty. So, she attempts to figure out what self-interest requires by investigating what honesty requires. If she succeeds, her motivating reasons will involve her knowledge that a particular action is honest, and therefore promotes self-interest. Since in every nearby possibility where the agent acts on that motive is a possibility where she performs the right action, her action would be safe from wrongness if we included that motivating reason in the basis of her action. But, clearly, we don't want to say that the agent's action is safe from wrongness even in this scenario: the motive of self-interest could easily have led her astray insofar as she could easily have found herself in a case where the de-

[11]This problem was first discussed by Goldman [1979] and developed later by Conee and Feldman [1998] in relation to process reliabilism.

mands of honesty diverge from the demands of self-interest. Anyone who appeals to the motivational basis of an action faces the burden of saying how coarsely we should individuate the basis of an agent's action.[12]

In response, I want to point out that the generality problem is only a *problem* for theorists who have the ambition of giving a precise theory of what makes a belief reliably true or of what makes an action reliably right. In response to this problem, therefore, we may simply give up this ambition: insofar as we do have a pre-theoretic grip over the notion of *safety from error* or *safety from wrongness* as it is applied in reasoning about knowledge and moral worth, it might well be permissible for us to use these notions in the course of theorizing without precisifying them any further.[13]

2.3 Against Reliabilism

MORAL RELIABILISM is an implausible hypothesis about moral worth. My hope, however, is that its defects will be instructive: they will tell us why EPISTEMIC RELIABILISM is also untenable.

2.3.1 Correctness Luck

The kind of luck that the reliabilist takes as her target in the moral as well as epistemic domains can be characterized quite generally. In Kant's example, the shopkeeper is lucky to perform a (subjectively) *right* action rather than a wrong one.

[12]One way to answering the question might be to restrict the basis of an action just to the *non-instrumental* motivating reasons of the action, i.e., to considerations about ends that the agent values non-instrumentally. I am worried that this might be too coarse a way of individuating the basis of an action. Imagine a scenario where I am motivated non-instrumentally by my concern for my friend's well-being to perform a certain action. But this could still lead me to perform a wrong action, say, in the case where I kill my friend's oppressive and violent husband for the sake of her well-being.

[13]For discussion, see Williamson [2000, p. 100; 2009, pp. 9-10]. Williamson, in fact, wants to claim that the imprecision of 'safety' must match the imprecision of 'knows.' I don't see why we should agree with that: unless we are given independent reason to think that safety from error is both necessary and sufficient for knowledge, there is no reason to think that there has to be any such match.

That is why his action lacks moral worth. The worth-undermining luck in this scenario pertains to the rightness of the action. Similarly, in *Fake Sheep*, Nina is lucky to form a *true* belief, rather than a false one. The knowledge-destroying luck in this case affects the truth of the belief. Now, moral rightness is a standard of correctness for actions, while truth is a standard of correctness for beliefs. Let us therefore call the relevant kind of luck *correctness luck.*[14]

CORRECTNESS LUCK.

1. An *action* suffers from *correctness luck* iff it is lucky to be right rather than wrong.
2. A *belief* suffers from *correctness luck* iff it is lucky to be true rather than false.

Many writers have recognized that a reliability condition like SAFETY FROM WRONGNESS cannot rule out all cases of correctness luck in the moral domain.[15] Imagine a variant of Kant's shopkeeper example, where an *invisible hand* ensures that every action performed from self-interest is morally right.[16] In such a world, the motive of self-interest reliably leads to morally right actions. So, the shopkeeper's actions are safe from wrongness. But he still seems lucky to have performed the right action, rather than the wrong one. Nomy Arpaly [2003] puts the point as follows:

[14]In the epistemic domain, correctness luck is labelled *veritic luck* by writers like Engel [1992] and Pritchard [2005].

[15]See, for example, Herman [1981], Stratton-Lake [2000], Arpaly [2003] and Markovits [2010].

[16]In his *Theory of Moral Sentiments*, Adam Smith Smith [2002/1759] comes close to taking such a world to be actual: "The proud and unfeeling landlord views his extensive fields, and without a thought for the wants of his brethren, in imagination consumes himself the whole harvest ... [Yet] the capacity of his stomach bears no proportion to the immensity of his desires ... the rest he will be obliged to distribute among those, who prepare, in the nicest manner, that little which he himself makes use of, among those who fit up the palace in which this little is to be consumed, among those who provide and keep in order all the different baubles and trinkets which are employed in the economy of greatness; all of whom thus derive from his luxury and caprice, that share of the necessaries of life, which they would in vain have expected from his humanity or his justice. The rich...are led by an invisible hand to make nearly the same distribution of the necessaries of life, which would have been made, had the earth been divided into equal portions among all its inhabitants, and thus without intending it, without knowing it, advance the interest of the society" (p. 215).

> [W]hat, exactly, is this "accidental" quality that we perceive in the grocer's doing of the right thing? We can, with some difficulty, imagine a world in which some invisible hand or other makes it true that the profit motive reliably produces morally right actions, and we can place Kant's grocer in that world, and still we shall not free ourselves from the sense that there is something accidental in the fact that he does the right thing. It is accidental in the same way as it is accidental that a person who reads *Lolita* for the love of scandal reads an aesthetically superior book, or the fact that a person who buys cheap beer because he likes it accidentally makes a money-saving choice. The former is attracted to the novel for reasons that are of no interest to the aesthetician who pronounces it beautiful, the latter is attracted to cheap beer for reasons that are of no interest to the thrifty, and Kant's grocer is attracted to fair pricing for reasons that are of no interest to the ethicist. The salient feature of Kant's case, I would like to suggest, is that the grocer's morally right action does not stem from any responsiveness on his part to moral reasons. (pp. 71-72)

Despite the reliability of self-interest with respect to moral rightness in these circumstances, the shopkeeper seems lucky to have performed the morally right action.

Even though the "invisible hand" variant of Kant's example brings out this point, there are more mundane examples of the same phenomenon. One function of social institutions, we might think, just is to make people reliably perform right actions on the basis of non-moral motives. For example, there might be a law that requires any person, present at the scene of a crime, to report that crime to an appropriate law enforcement official (to the extent that doing so doesn't threaten his or her personal safety). If the public knows that there is such a law, then many people, simply in fear of penalty and not out of concern for the victim of the crime, might report crimes to law enforcement officials. In doing so, they would be performing the morally right action. Moreover, if the lawmaking bodies in this community are sufficiently just, then the motive that they would be acting on couldn't easily lead them astray in the relevant circumstances. But, given the manner in which these agents act, their actions still seem lucky to be right; for the motive behind these actions does not have anything to do with the reason why these actions are right.

Here is a diagnosis of this intuition. In these scenarios, there is a modally robust connection between the basis of the agent's action and moral rightness; that makes her act reliably well. But the connection is forged by a feature of the agent's predicament—the presence of the invisible hand or of a just legislature—which has nothing to do with her motivational profile. That is why, in the light of her motivational profile, it still seems like a matter of luck that she performs the right action. From this, it follows that moral reliability isn't sufficient to block the kind of luck that destroys moral worth.

A similar complaint can be raised against EPISTEMIC RELIABILISM. Gettier cases like *Fake Sheep* are also instances of *correctness luck*: in such scenarios, the agent's belief is lucky to be true, rather than false. Even though SAFETY FROM ERROR rules out some instances of correctness luck like *Fake Sheep*, it isn't sufficient for blocking such luck altogether. To see why, consider the following variation on *Fake Sheep*.

> *Fake Sheep Redux.* Nina is looking at a field, wondering whether there is a sheep out there. At that time, an object that looks like a sheep comes into view. Thinking it's a sheep, Nina comes to believe that there is a sheep in the field. However, what Nina sees is not a sheep, but a sheepdog camouflaged as one! Luckily, Nina's belief is true: there is a sheep in a different part of the field which she cannot see. In fact, Nina is in 'hardworking sheep-dog country', where a sheepdog is never seen in a field unless it is keeping watch over a sheep in the same field. Still, it seems that Nina lacks knowledge.[17]

In *Fake Sheep Redux*, Nina is in a region where she couldn't easily have fallen into error in forming a belief in the same manner as in the actual scenario. Intuitively, at least, her belief is safe from error. Yet, it seems lucky that she hit the truth rather than a falsehood, given the manner in which she forms her belief.

The analogy between knowledge and moral worth suggests a diagnosis of this intuition. In the "invisible hand" version of Kant's shopkeeper example, a feature

[17] This example is borrowed from Miracchi [2015].

of the agent's predicament that reflects nothing about his motivational profile—namely, that he lives in a world where the demands of honesty coincide with the demands of self-interest—makes his action safe from wrongness. In *Fake Sheep Redux*, a feature of the agent's environment that has nothing to do with the basis of her belief—namely, the fact that she is in hard-working sheepdog country—seems to make it safe from error. For that reason, given the facts about the basis of her belief, the agent still seems to lucky to have formed a true belief.

The challenge generalizes. Any reliability condition on knowledge is concerned with the truth-conduciveness of the basis of a belief. So, it will inevitably impose a constraint on the modal profile of the relevant basis: it will require the basis of the belief to yield true beliefs in all nearby cases. However, such a modally robust connection between the basis of the belief and its truth could be induced by features of the agent's predicament which have nothing to do with the basis of the agent's belief. Such features are therefore insignificant for the purposes of evaluating the manner in which the agent forms her beliefs. That is why, given the facts about the basis of her belief, the agent would still seem lucky to have formed a true belief. Thus, truth-conduciveness of the basis of a belief isn't sufficient to block correctness luck.

A reliabilist might push back against this challenge by insisting that the notions of *safety* and *reliability* are merely formal tools, which we use in our theorizing and which are constrained only by our pre-theoretic judgments about knowledge.[18] So, given that Nina's belief in *Fake Sheep Redux* intuitively doesn't count as knowledge, we might insist that there is indeed a sufficiently similar case where she forms a belief that there is a sheep in the field on the basis of seeing the sheepdog, but there is no sheep around, making her belief false. I am not sympathetic to this approach.

It seems to me that even the reliabilist should admit that we have a pre-theoretic grip on the notion of 'safety' or 'reliability': unless she does so, she won't be able to explain why there is intuitive pressure to accept safety as a necessary condition on knowledge. Presumably, the manner in which the reliabilist defends her view

[18]See, for example, Williamson's [2009, 2009] response to putative cases of unsafe knowledge discussed by Neta and Rohrbaugh [2004].

involves an inference to the best explanation from pre-theoretic judgements about cases where the absence of reliability or safety results in the absence of knowledge. Once the reliabilist accepts that we do have an intuitive grip on what belief counts as safe and what doesn't, we could indeed argue by appeal to cases like *Fake Sheep Redux* that there are beliefs that are intuitively safe, but do not count as knowledge.

2.3.2 Defeat Luck

Correctness luck isn't the only form of luck that undermines knowledge and moral worth. There is another kind, which I shall call *defeat luck*.

DEFEAT LUCK.

1. An *action* suffers from *defeat luck* iff it is a matter of luck that the agent's reasons for performing the action are undefeated by morally suboptimal features of her character.
2. A *belief* suffers from *defeat luck* iff it is a matter of luck that the agent's reasons for holding the belief are undefeated by misleading evidence.

To see what defeat luck consists in, consider the following scenario discussed by Arpaly [2003].

> [I]magine the person who acts benevolently on a whim. It is Sunday morning and she is awakened by a call from a charity asking for a donation. Our agent thinks, "Why not do something right?" and is moved to do something right so long as her credit card happens to be close enough to the bed. [This] agent—the person whose moral concern is skin deep—would be very presumptuous to expect much praise for an action that almost seems accidental, attributable to the charity's call and the location of the credit card more than to her depth of concern for her fellow human beings. Still, there is no reason to doubt that she has acted for moral reasons. When a person whimsically asks for milk instead of cream in the coffee she has with her chocolate cake, one does not doubt that she does it for health reasons but doubts merely the seriousness of her concern. (pp. 87-88)

Call this agent the *capricious philanthropist*. Intuitively, her action lacks moral worth, because she is lucky to perform the right action. Can a reliability condition like SAFETY FROM WRONGNESS block such luck? It can't: in all the nearby cases where the capricious philanthropist acts on the same basis as in the actual scenario—intuitively, on a motive of benevolence—her action is morally right. What then makes her success at performing the right action lucky in this case? Even though there may not be nearby cases where the philanthropist's benevolent motive generates the wrong action, there indeed are nearby cases where it doesn't generate any action. If her wallet had been a few feet further away from her bed, she would have noticed it, and wouldn't have acted on her benevolent motive. In the actual scenario, therefore, it is a matter of luck that the capricious philanthropist's reasons for performing her action remain undefeated by her reluctance to get up from bed. This kind of luck—which a reliability condition cannot exclude but which nevertheless undermines moral worth—is what I shall call *defeat luck*.

Are there analogous cases of defeat luck in the epistemic domain? Arguably, there are. Gilbert Harman [1973] describes some examples where misleading evidence that the agent does not possess undermines the claim of her belief to knowledge.

> *Assassination.* A political leader is assassinated. An enterprising and reliable journalist witnesses the assassination and writes a report on it, which is then printed in the final edition of a newspaper. Jill reads the newspaper report, and comes to believe that the political leader is dead. But later the associates of the political leader, fearing a coup, decide to pretend that the bullet hit someone else. On nationwide television they announce that an assassination attempt has failed to kill the leader but has killed a secret service man by mistake. Rationally trusting this announcement, most people come to abandon their earlier beliefs about the death of the political leader. Luckily, Jill wasn't watching television when this announcement was broadcast. She continues to believe that the leader was assassinated.

Tom and Buck. I am the library detective. One day, I see Tom smuggling a book out of the university library under his coat, and thus come to know this. I then testify before the University Judicial Council, saying that I saw Tom stealing a book from the library. After testifying, I leave the hearing room. Later that day, Tom's mother testifies at the same hearing. She claims that Tom couldn't have committed the theft; he was thousands of miles away at the time of the theft. However, she says, Tom's identical twin, Buck, was around, and he is an inveterate kleptomaniac.

For Harman, in each of these cases, a piece of evidence that the agent doesn't possess makes the agent lose her knowledge. In *Assassination,* surrounded by people who have evidence contrary to what she believes, Jill fails to retain her knowledge that the political leader is dead. In *Tom and Buck* everyone at the hearing rationally believes that Tom didn't steal the book after Tom's mother testifies. The wide availability of information about Tom's mother's testimony destroys my knowledge that Tom stole the book.

Many writers have taken Harman's judgements at face value. For Keith Lehrer [1974, pp. 221-223] and Peter Klein [1981, §§3.9-10], these cases provide evidence for a *defeasibility* theory of knowledge, under which a belief counts as knowledge only if there is no misleading defeater that the agent is unaware of. Robert Nozick [1981, p. 177] has used them to motivate his *adherence* condition on knowledge, i.e., the condition that a belief that is knowledge must be held by the agent in all the close possibilities where the content of the belief is true and the agent arrives at the belief by the same method as in the actual scenario. More recently, Timothy Williamson [2000] has taken Harman's cases to suggest that "present knowledge is less vulnerable than merely true belief to *rational* undermining by future evidence" (p. 79).[19]

[19]Other writers have rejected Harman's intuitions about these cases. However, the considerations they have offered against them do not seem particularly strong. For example, Lycan [1977] thinks that Harman's intuitions about these cases arise from a version of the *requirement of total evidence* that requires an agent to base her beliefs on all the evidence that is easily accessible to her, both

If Harman and these writers are right, then we need to explain why *Assassination* and *Tom and Buck* are not instances of knowledge. At first glance, these look like cases of epistemic luck; it is a matter of luck that the agent believes the truth in each of these cases. In *Assassination,* Jill could easily become aware of the television broadcast and lose her belief about the death of the political leader. Similarly, in *Tom and Buck,* after Tom's mother testifies, I could easily come to know about her testimony and lose my belief that Tom stole the book. Hence, we may presume that these cases instantiate some form of epistemic luck that is incompatible with knowledge.

What does *this* kind of knowledge-destroying epistemic luck consist in? Intuitively, these are not cases of correctness luck: there is no reason to think that the agent's belief is lucky to be true in any of these cases. In *Assassination,* the newspaper report that forms the basis of Jill's belief is generated by a reporter who witnessed the political leader's death. So, it is no accident that her belief is true rather than false. Similarly, in *Tom and Buck,* provided that Tom's mother is lying and Tom has no twin, my capacity to recognize people might be working so well that it is no accident that I correctly, rather than incorrectly, identified Tom when I saw him. For this reason, SAFETY FROM ERROR is unsuitable for ruling these cases out.

---

possessed and unpossessed. Lycan takes this requirement to be implausible and therefore rejects these intuitions. By contrast, Engel [1992] treats Harman's cases as instances of *evidential epistemic luck,* where the agent is merely lucky to have the kind of evidence that she has. Since such luck isn't incompatible with knowledge, he is willing to concede that the relevant agent does know in these cases. The dialectical strategy adopted by Lycan and Engel is roughly the same: they attempt to find what they think is the most plausible diagnosis of Harman's intuitions and then show that the relevant diagnosis does not withstand critical scrutiny. If a more defensible diagnosis of these intuitions could be found, their arguments would have no force. I intend to offer such a diagnosis.

Some other arguments against Harman's intuitions conflict with more widely endorsed intuitive judgements. For instance, Hetherington [1998] takes these examples to be scenarios where the agent could easily have lost her knowledge, but actually knows the relevant claim. This diagnosis leads him to say that many traditionally accepted Gettier cases, like Goldman's [1976] fake barn country example, are in fact instances of knowledge. The same approach can be seen in Lycan [?], who takes knowledge to be justified true belief not based on false premises. In Harman's cases as well as in the fake barns case, the agent has such a belief. So, Lycan is committed to ascribing knowledge in both these cases. This seems like a theoretical cost to me.

Later in this paper, in Section 6, I say why there might be a conflict of intuitions about these examples.

As long as Jill and I form our beliefs on the same bases as we do in *Assassination* and *Tom and Buck* respectively, there may be no nearby cases where we form a false belief on the same bases. So, our beliefs are free from the risk of error.[20]

A better diagnosis is that, in Harman's cases, there is an easily accessible fact about the agent's predicament, which, if discovered by her, would make her lose her reason for believing what in fact is true. In *Assassination*, it's the television announcement. In *Tom and Buck*, it's the testimony given by Tom's mother. So, in these cases, it is a matter of luck that the agent's belief isn't rationally defeated by misleading evidence.[21] That is why the agent seems to be in an epistemically precarious state. This kind of luck, which is different from correctness luck, is what I have called *defeat luck*.[22]

The insight, therefore, is this: both knowledge and moral worth are undermined

[20]Some have attempted to reduce the relevant kind of luck to correctness luck. In relation to *Assassination*, for example, Pritchard [2005, Chapter 6] seems to suggest that our hesitation to ascribe knowledge to Jill in this case can be explained by the fact that we don't in fact take Jill's belief to be free from risk of error. Since she lives in a state where the state interferes with the media, the belief that Jill forms on the basis of the newspaper report isn't reliable. It is not obvious to me that the state interference is an essential part of the description of the case. We could imagine a similar scenario where the public denial of the leader's death was a one-time prank that an otherwise globally reliable news channel chose to play on its viewers.

Another diagnosis of Harman's intuitions could be that in these cases, the protagonist is simply lucky to have the evidence that she in fact has. Some writers like Engel [1992] and Pritchard [2005, Chapter 6] take this to be the correct diagnosis of Harman's intuitions. For a detailed discussion of evidential epistemic luck, see Pritchard [2005, Chapter 5]. As pointed out before, this is what is sometimes called *evidential luck*. However, this cannot be the right diagnosis. For evidential luck isn't incompatible with knowledge. Nozick [1981, p. 193] describes a scenario where a masked bank-robber is escaping the crime scene, when his mask slips and a bystander by happenstance sees that it is Jesse James. Surely, in examples of this sort, the agent does know. So, even an evidentially lucky belief can be knowledge.

[21]As I understand *rational defeat*, the evidential support for a belief in a proposition P is *rationally defeated* at time t if and only if the agent's total evidence changes from E to E' at time t due to the rational impact of new evidence, such that the evidential support that E provides to P is not provided by E'. Under this construal, the evidential support for a belief cannot be rationally defeated by a fact, or a true proposition, unless the agent acquires that fact as new evidence.

[22]What I am calling *defeat luck* need not be completely distinct from evidential luck, but may just be a variety of evidential luck. A belief is evidentially lucky if and only if the agent is lucky to have the evidence she actually does. If an agent is lucky not to have encountered defeating evidence against a belief of hers, then her belief may be subject to evidential luck of some sort. Even if defeat luck is a form of evidential luck, this still doesn't mean that other forms of evidential luck are incompatible with knowledge, or that defeat luck is compatible with knowledge.

by defeat luck. Not only does this insight reveal another respect in which knowledge resembles moral worth, but it also shows why knowledge is a better analogue of moral worth than doxastic justification. Intuitively, it seems that, for a belief to be doxastically justified, it is enough that the agent's evidence adequately supports the content of the belief, and she forms and maintains her belief on the basis of that evidence. So, it is possible for an agent to be in a position where her reasons for holding a belief could easily have been defeated by epistemically suboptimal feature of her predicament (not just by misleading evidence, but also by other kinds of distraction), and still be doxastically justified in holding her belief. Thus, nothing analogous to defeat luck obviously undermines doxastic justification. This is an important disanalogy between moral worth and doxastic justification.

Here is the upshot. We have seen that there are two kinds of luck that undermine knowledge and moral worth: correctness luck and defeat luck. A reliability condition like SAFETY FROM ERROR isn't sufficient to block either.

2.4 Moral Explanationism

Some writers, like Nomy Arpaly and Julia Markovits, argue that if a right action is free from correctness luck—i.e., if it isn't just a matter of luck that she performs the right (rather than the wrong) action, then the motivating reasons for the action must coincide with the facts that make the relevant action morally right. Arpaly [2003] puts the point as follows.

> In pricing fairly, the grocer acts for a reason that has nothing to do with morality or with the features of his action that make it morally right. The reasons for which he acts have to do only with his own welfare; and whatever it is that makes his action morally right, the fact that his action increases his welfare is certainly not what makes it morally right. His reasons for action do not *correspond* to the action's right-making features. (p. 72)

The facts that make an action morally right are the *normative reasons* for that action.

These are the facts which explain why an action is right.[23] The view that Arpaly lays down above, therefore, is that, for an action to have moral worth, the motivating reasons for the action must coincide with the normative reasons for her action, i.e., the facts that explain why the action action is right. Taking this condition to be both necessary and sufficient for moral worth, Julia Markovits [2010] writes:

> According to what I will call the Coincident Reasons Thesis, *my action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action*—that is, if and only if I perform the action I morally ought to perform, for the (normative) reasons why it morally ought to be performed. (p. 205)[24]

I am going to accept only a part of Markovits' thesis: namely, that coincidence between the normative reasons for an action and its motivating reasons is necessary for it to have moral worth. This explains our intuitive judgement about Kant's shopkeeper. In the scenario where the shopkeeper performs a right action from self-interest, what makes the action right for the shopkeeper to perform is the fact that, by his lights, the relevant price is the fair one. However, what motivates the agent is the belief that the action will result in profit. Since the motivating reasons for the action do not explain why the shopkeeper performs a right action rather than a wrong one in the relevant circumstances, he seems lucky to have performed the right, rather than the wrong, action. That is why his action lacks moral worth.

The more general insight that emerges from this is that an action suffers from correctness luck if and only if the motivating reasons underlying the action fail to explain why the agent performs a *right* action rather than a wrong one in those

[23]For a defence of this view, see John Broome [2004, 2013]. Broome [2013, p. 50] writes: "A pro toto *reason* for N to F is an explanation of why N ought to F."

[24]Stratton-Lake [2000] ascribes this thesis, under the label "symmetry thesis" to Korsgaard [?] who claims that, for Kant, "*the reason why a good-willed person does an action, and the reason why the action is right, are the same*" (italics in the original, p. 60). Hursthouse [1999] also defends a similar view. For Stratton-Lake, the coincidence of motivating reasons and normative reasons is neither necessary, nor sufficient for moral worth. For Arpaly [2003], it is necessary, but not sufficient for moral worth. The version of the thesis, to which Markovits [2010] subscribes, makes the coincidence between motivating reasons and rightmaking reasons not only necessary, but also sufficient for moral worth.

circumstances.[25] Call this the RIGHTNESS REQUIREMENT.

> RIGHTNESS REQUIREMENT. A right action is free from correctness luck if and only if the basis of the action explains why the agent performs a *right* action rather than a *wrong* one.

What about defeat luck? The capricious philanthropist isn't like Kant's shopkeeper: she does act for the reasons which make her action morally right. Arpaly's [2003] discussion suggests that cases of this kind could be handled by requiring that an agent's action has moral worth only if the agent exhibits a certain amount of *depth of moral concern.* Under one interpretation, this means that for an agent's action to have moral worth, not only must she act for the rightmaking reasons, but she also should have a secondary dispositional motive—e.g., the motive of duty in the Kantian framework or a character trait in an Aristotelian framework—which makes her act *somewhat stably* on the basis of rightmaking reasons across a range of counterfactual scenarios.[26]

[25]The kind of explanation that is relevant to such non-accidental success is therefore *contrastive.* For the view that explanations are essentially *contrastive,* see Bromberger [1992, Chapter 3], van Fraassen [1980] and Garfinkel [1981]. For my purposes, it is just enough to say that some explanations, but not all, are contrastive.

[26]One might worry that this isn't very informative. Whatever secondary dispositional motive may be required for an action to have moral worth, it need not be invulnerable to all morally neutral or vicious psychological influences; for an agent like us, there is always a breaking point where good motives would be defeated by some morally suboptimal feature of her psychology. Julia Markovits [2010], for example, describes the case of a fanatical dog-lover who saves the lives of some strangers at great risk to himself. If his dog had been in danger, the dog-lover would know it and would try to save the dog rather than the strangers. Even if he felt some concern for the strangers, that concern would be overcome by his excessive concern for the dog. Now, suppose that his dog could easily have been in danger. Does the nearby possibility of such motivational defeat undermine the moral worth of the dog-lover's action? It seems not. So, the challenge for someone who takes a secondary dispositional motive to be necessary for moral worth is to say what sorts of potentially defeating factors could undermine the worth of an agent's actions.

I want to reject this challenge. It seems that the dog-lover, under normal circumstances, is a person who cares about the well-being of other people. But as soon as his dog is in danger, he is only concerned about the well-being of his dog. Thus, his patterns of motivation when the dog is in danger seem like a big departure from the normal motivational profile which guides his actual action. That is why, pre-theoretically, at least, whether or not the dog was in danger doesn't seem relevant to evaluating his actual action. By contrast, whether the capricious philanthropist takes her wallet to be too far away from her bed does seem to matter. The capricious philanthropist, even

To my mind, these second-order motives are just background conditions that enable the actual basis of agent's action to explain why she performs the relevant action rather than refraining from performing it. Absent such motives, the actual motivating reasons for the agent's action wouldn't be able to play this explanatory role; for there would be a salient alternative possibility where the agent is motivated to act on the same basis, but her good motives are defeated by some distraction. Take the case of the capricious philanthropist. We cannot explain why she performs the action rather than refraining from performing it, simply by appeal to the motivating reasons behind her action. There is a salient alternative possibility where she is motivated by these reasons, but still fails to perform the benevolent action because she sees that her wallet is too far away from her bed. Thus, in order to explain why she performs her action rather than refraining from performing it, we also have to invoke the fact that she didn't think that her wallet was far away from her bed. Since this latter fact isn't included amongst the motivating reasons for her action, the basis of her action—which consists of those motivating reasons—will fail to explain why she performs the action, rather than not performing it.

So, we might think: for an agent's action to be free from defeat luck, the motivating reasons underlying her action must explain why the action is performed rather than omitted. Call this the PERFORMANCE REQUIREMENT.

> PERFORMANCE REQUIREMENT. A right action is free from defeat luck if and only if the basis of the action explains why the agent *performs* an action rather than *refraining* from performing it.

Assuming that correctness luck and defeat luck are the only forms of luck that undermine moral worth, we obtain the following anti-luck condition on moral worth.

---

under normal circumstances, seems to be subject to easy distractions. That is why the manner in which she is motivated when she discovers that her wallet is not very close to her bed isn't a big departure from her actual motivational profile; so, what happens in that case does seem relevant to the assessment of the manner in which the agent acts. What counts as a big departure from an agent's actual motivational profile cannot be specified precisely; all we can rely on here are our intuitive judgements about particular cases. The concept of *moral praiseworthiness* or *positive moral worth* is essentially vague; if we want our theory of moral worth to be faithful to that vagueness, then we should resist the temptation of laying out in general terms what sorts of potentially defeating factors could make an action unworthy of moral praise.

MORAL EXPLANATIONISM. A right action is free from luck that undermines moral worth if and only if the basis of the action explain both why the agent performs a *right* action rather a *wrong* one, and why the agent *performs* the relevant action, rather than *refraining* from performing it.

2.5 EPISTEMIC EXPLANATIONISM

Can the explanationist conception of worth-undermining moral luck generalized to cover the case of knowledge-destroying epistemic luck? I want to argue that it can be; for it falls out of a more general conception of *accident.*

Suppose tomorrow is my wedding day and I pray that it doesn't rain. It doesn't rain. The faithful will insist that this was no accident; God heard my prayer and answered it. The sceptic will deny this. She will say that the fine weather on my wedding day is explanatorily independent of my prayer; there is no explanation, divine or otherwise, of why these events coincided. This fits an old suggestion of Aristotle's: an event counts as an *accident* just in case there is no explanation of why all the constituents of the event *coincide.*[27,28]

This of course isn't to say that there is no explanation of why my prayer came to true. After all, there might be an explanation of why I prayed—i.e., my desire to get married in fine weather—*and* an explanation of the fine weather—i.e., facts

[27]I borrow this example and definition of *accident* from Owens [1992] who discusses the same phenomenon under the label 'coincidence.' In *Physics* (II 4-6) and *Metaphysics* (V 30, VI 3, XI 8), Aristotle presents us with a catalogue of accidental conjunctions: going to the well to drink water after eating some spicy food and meeting some ruffians who are passing by, finding a buried treasure while digging a hole for a plant, meeting one's debtor while at the marketplace on other business, being a pastrycook and curing someone, etc. We might wonder: What is it that makes these conjunctions of events *accidents*? The answer that Aristotle defends is that these conjunctions have no causes. Suppose I go to the well to drink water after eating some spicy food and meet some ruffians who are passing by. The ruffians then kill me. In that case, Aristotle would say, the cause of my death would be being at the well *at the same time* as the ruffians. But this event, Aristotle claims in *Metaphysics* VI 3 1027b12-14, has no cause. For discussion, see Sorabji [1980].

[28]Here, I am using the word 'event' quite loosely, as synonymous with 'fact'. What then are consituents of an event? In my opinion, these also are facts, which stand in some parthood relation to more complex facts.

about air currents in the relevant region. Conjoining these explanations might indeed give us all the information we need to explain why my prayer came true. Still, given that these explanations are completely unconnected, they still wouldn't explain why my prayer *coincided* with the fine weather the next day. That is what makes the event an *accident*. Let us apply this *explanation*-based conception of accident to the case of moral worth.

When we are evaluating an action for moral worth, we are asking the following question: given the manner in which the agent performs the action, is her success at doing the right thing an accident? Note that the agent's success at doing the right thing could be understood as constituted by two elements: the fact that she performs the action and the fact that the action turns out to be right. According to the explanation-based account of accident, therefore, an agent's success at doing the right can be an accident in the light of the basis of her action just in case the basis fails to explain why these two facts coincide. Now, there are two ways in which that could happen. On the one hand, the basis might explain why the agent performs the action rather than refraining from performing it, but might fail to explain why the she acts rightly rather than wrongly. On the other hand, it might explain why the agent acts rightly rather than wrongly, but might fail to explain why the agent performs the action rather than refraining from performing it. So, for an agent's action to have moral worth, her motivating reasons must *explain both* why she performs a right action rather than a wrong one, and why she performs the relevant action rather than refraining from performing it. This is precisely what MORAL EXPLANATIONISM requires.

We can extract a similar account in the epistemic case. When we are assessing whether a belief amounts to knowledge, we are asking the following question: given the manner in which the agent forms her belief, is her success at believing the truth an accident? Here, again, the agent's success at believing the truth could be understood as constituted by two elements: the fact that the agent holds the relevant belief and the fact that the belief happens to be a true one. So, an agent's success at believing the truth can be a coincidence just in case the basis of her belief fails to explain why these two facts obtain together. The basis of a belief might fail

to play this explanatory role for two reasons. On the one hand, it might fail why the agent possesses the relevant belief rather than lacking it, but might fail to explain why the agent holds a true belief rather than a false one. On the other hand, it might explain why the agent comes by a true belief rather than false one, but might still fail to explain why the agent possesses the belief rather than lacking it. Hence, for an agent's belief to be knowledge, the basis of her belief must explain why she comes by a true belief rather than a false one, and why she possesses the relevant belief rather than lacking it.

> EPISTEMIC EXPLANATIONISM. A belief is free from knowledge-destroying epistemic luck if and only if the basis of the belief explains both why the agent holds a *true* belief rather than a *false* one, and why the agent *possesses* the relevant belief rather than *lacking* it.

Just as MORAL EXPLANATIONISM imposes two distinct conditions to block correctness luck and defeat luck respectively in the practical domain, so also does EPISTEMIC EXPLANATIONISM impose two distinct conditions to handles these two forms of luck in the epistemic domain.

> TRUTH REQUIREMENT. A true belief is free from correctness luck if and only if the basis of the belief explains why the agent holds a *true* belief, rather than a *false* one.
>
> POSSESSION REQUIREMENT. A true belief is free from defeat luck if and only if the basis of the belief explains why the agent *possesses* the relevant belief, rather than *lacking* it.

According to the TRUTH REQUIREMENT, what blocks correctness luck in the epistemic case is the the right sort of explanatory relationship between the basis of a belief and the fact that the agent holds a true belief rather than a false one. In *Fake Sheep*, the basis of Nina's belief doesn't explain why the agent holds a belief that is true rather than false. The experience on the basis of which Nina forms her belief is generated by perceptual interaction with a dog camouflaged as a sheep. What explains the truth of Nina's belief is that there is a sheep in the field. Since this

isn't part of the basis of his belief, her belief doesn't satisfy the TRUTH REQUIREMENT of explanationism. Similarly, in *Fake Sheep Redux*, we may assume that the agent forms her belief in the same manner as in *Fake Sheep*: by looking at the camouflaged dog. In this example, what explains the truth of Nina's belief could not just be the fact that she looked at a camouflaged dog, but also the fact that the dog wouldn't be present unless there was a sheep in the same field. Since this latter fact doesn't enter into the basis of the agent's belief, the basis cannot explain why the agent forms a true belief rather than a false one.[29]

According to the POSSESSION REQUIREMENT, defeat luck obtains just in case the basis of the agent's belief fails to explain why her belief is held rather than lost. For example, Jill's belief in *Assassination*, is based on a newspaper report given by a reporter who was a witness of the leader's assassination. What explains why she holds her belief is the causal mechanism that forms the basis of her belief *plus* the fact that she wasn't watching television when the announcement was broadcast. This last fact definitely isn't part of the basis of her belief. So, Jill's belief doesn't satisfy the POSSESSION REQUIREMENT of explanationism. Therefore, the POSSESSION REQUIREMENT rules out this case. Similarly, in *Tom and Buck*, the basis of my belief may indeed explain why my belief is true rather than false; after all, I formed my belief after seeing Tom steal the book from the library. However, given the availability of evidence about Tom's mother's testimony, the basis of my belief cannot explain why I possess my belief rather than lacking it. In order to explain this, we would have to appeal to the fact that I wasn't present at the hearing when

[29]One might argue that the basis of the agent's belief in *Fake Sheep Redux* isn't quite the same as in *Fake Sheep*. "After all," one might say, "the fact that the agent is in hardworking sheepdog country is an epistemically significant factor of her environment on which the formation of her belief causally depends; so, it should included in the basis of her belief. Once we do that, the basis of her belief will indeed explain why she comes by a true belief rather a false one." It is not clear to me that the agent's belief causally depends on the fact that the agent is in hardworking sheepdog country; after all, even if the agent weren't in hardworking sheepdog country but saw a camouflaged sheepdog at a distance, she would still have come to believe that there is a sheep in the field. In this sense, the agent's formation of the belief isn't counterfactually dependent on being in hardworking sheepdog country. If we take counterfactual dependence to be necessary for causal dependence, then the formation of her belief cannot causally depend on the fact that she is in hardworking sheepdog country.

she testified. This fact definitely isn't part of the basis of my belief. Once again, the POSSESSION REQUIREMENT predicts that this isn't a case of knowledge.

2.6 Advantages of Epistemic Explanationism

Let me highlight some advantages of adopting this explanation-based conception of knowledge.

2.6.1 Other Explanation-Based Accounts of Knowledge

On the Arpaly-Markovits picture of moral worth, an agent's action has moral worth only if the fact makes the agent's action right explains why the agent performs the relevant action. An exactly analogous view about knowledge should say that a belief amounts to knowledge only if the fact that makes the agent's belief true explains why the agent holds the relevant belief. Such a view is advocated by Goldman [1988] and Jenkins [2006].

I find this hypothesis about knowledge quite appealing. This appears to be true about perceptual knowledge, where the agent seems to come into direct cognitive contact with the external world. Some philosophers of perception, for example, have claimed that what is special about veridical perception that it reveals to the agent the truthmaker of the content that she perceptually believes.[30] The hypothesis also fits slightly more mediate forms of knowledge, like inferential and testimonial knowledge. If a detective finds a fingerprint on the murder weapon, and the fingerprint turns to be that of the butler, she may thereby come to know that the bulter committed the murder. Since the detective's evidence is explained by the fact that makes the detective's belief true, the fact that her belief concerns explains why she holds her belief. Or, if you sincerely tell me that you cooked pasta today, I could thereby come to know this. Here, what makes my belief true is the fact that you cooked pasta. If this explains why you told me that you cooked pasta, then the fact makes my belief true can explain why it is held.

[30]For example, see Johnston [2006] and Fish [2009].

However, the view runs into trouble when it is applied to, say, mathematical knowledge knowledge or knowledge of the future. If we have a Platonist conception of mathematical facts, for instance, we might argue that such facts are causally inert. If they are causally inert, they cannot possibly stand in any causal-explanatory relationship with our beliefs. Similarly, it is unclear whether facts that make beliefs about the future true—presumably, future facts—can enter into causal explanations of why a belief was held at a prior time. None of these worries, I must say, are decisive. We may simply take the plausibility of this theory of knowledge to be an argument against Platonism, and we may allow future facts to explain past beliefs. Since I am unwilling to take a stance on such issues, I want to propose that we frame our theory of knowledge in a more neutral manner.

EPISTEMIC EXPLANATIONISM allows us to do so. For an agent's belief to be knowledge, the fact makes the belief true need not causally explain it. However, the basis of the belief, which explains why the belief is held, must explain why the agent forms a true belief rather than a false one. In the case of mathematical knowledge, even if mathematical facts aren't part of the bases of our mathematical beliefs, those bases—as long as they involve the exercise of some capacity to reliably detect mathematical truths—may still be able to explain why we come by true mathematical beliefs rather than false ones. Similarly, in the case of knowledge about the future, as long as the basis of an agent's belief involves knowledge of the regularities that will eventually make certain future facts come out true, the basis may be able to explain why she forms a true belief rather than a false one. In this manner, the TRUTH REQUIREMENT avoids some of the potential difficulties that other explanation-based accounts of knowledge face.

2.6.2 Conflict of Intuitions

Many writers take defeat luck to be compatible with moral worth and knowledge. Markovits [2010], for instance, has argued that performing the right action for the right reasons is sufficient for moral worth. According to view, therefore, when the capricious philanthropist acts benevolently out of concern for other people,

her action is indeed morally praiseworthy. Similarly, Lycan [1977], Engel [1992], and Pritchard [2005] have argued against the hypothesis that defeat luck undermines knowledge. So, in Harman's assassination example, even though there is easily accessible misleading evidence in Jill's environment, she still gets to keep her knowledge. What motivates these arguments, I think, is a different set of intuitions about these cases: some are inclined to ascribe moral worth to the capricious philanthropist's action, and knowledge to Jill in Harman's assassination example. EPISTEMIC EXPLANATIONISM helps to see why such a conflict of intuitions might arise.

According to certain views of causation (e.g. the view on causation involves some kind of physical connection between events), *omissions* and *absences* cannot be causes: for example, the event of my currently not inhaling some lethal gas cannot intelligibly be regarded as a cause of my staying alive.[31] If this is right, then in Arpaly's capricious philanthropist case, the philanthropist's not taking her wallet to be too far away from the bed cannot cause her benevolent action. Similarly, in Harman's example, Jill's not watching the television announcement cannot cause her to possess the belief that the political leader is dead.

Now, we may put the following constraint on causal explanation: a fact that an event *e* obtains can be part of a causal explanation just in case *e* is an event that is capable of causing the event to be explained. In that case, the fact that the philanthropist didn't take her wallet to be too far away from her bed cannot be part of a causal explanation of why she performs the relevant action. Hence, the basis of her action might indeed be sufficient to causally explain why she performs the relevant action. Similarly, the fact that Jill didn't see the television announcement will not be part of the causal explanation of why she possesses the belief that political leader is death. In that case, the basis of Jill's belief may include all the facts that are required to causally explain why she possesses that belief. This might explain why some might take the case of the capricious philanthropist to be a case of moral

[31]For the case against causation by omission, see Aronson [1971], Dowe [2001, 2004], Armstrong [2001], and Beebee [2004]. For the case in favour of causation by omission, see Lewis [2004] and Schaffer [2000, 2004].

worth, and Harman's examples to be instances of knowledge.

However, it is not clear to me whether this line of reasoning survives reflection. Even though omissions and absences might not be causes, facts about such events could still play a role in (causal) explanations: for example, if there is a looming threat that I might be killed by a poisonous gas, then an explanation of why I stay alive might have to appeal to the fact that I did not breathe in any poisonous gas (or, at least, some negative fact in the vicinity).[32]

This yields the following conclusion about the capricious philanthropist. In that scenario, there is a salient alternative possibility where the philanthropist sees that her wallet is too far away from her bed, and doesn't donate money. So, in order to explain why she performed the benevolent action in the actual case rather than not performing it, we have to appeal to the fact that she didn't take her wallet to be too far away from her bed. So, the motivational basis of her action by itself will be insufficient for explaining why the action was performed rather than not performed. As a result, her action won't satisfy the performance requirement, and therefore won't have moral worth. Analogously, in Harman's assassination example, the alternative possibility where Jill loses the belief upon hearing the television announcement is salient. So, if we are asked to explain why Jill possesses the relevant belief rather than lacking it, we have to invoke the fact that she wasn't watching television when the announcement was made. Therefore, her belief will fail to satisfy the POSSESSION REQUIREMENT, and won't count as knowledge.

2.7 The Structure of Knowledge

We have seen that EPISTEMIC EXPLANATIONISM is able to handle cases of correctness luck and defeat luck adequately. Even though EPISTEMIC EXPLANATIONISM does seem to generate the right prediction about several instances of correctness luck and defeat luck, it would be a mistake to expect it to yield concrete predictions about every imaginable example of epistemic luck. Just like SAFETY FROM ERROR,

[32]Even opponents of causation by omission concede this point. See, for instance, Beebee [2004].

EPISTEMIC EXPLANATIONISM involves the essentially vague notion of *basis*, which might not be easy to specify in particularly tricky cases. Therefore, EPISTEMIC EXPLANATIONISM may not give us any general algorithm for telling whether any arbitrary belief suffers from knowledge-destroying epistemic luck.

This, however, need not make EPISTEMIC EXPLANATIONISM uninformative. EPISTEMIC EXPLANATIONISM illuminates various structural features of knowledge. It explains the appeal of *relevant alternatives* theories of knowledge. It also vindicates the thought that knowledge requires both reliability and stability.

2.7.1 Relevant Alternatives

Many writers have thought that knowledge is essentially contrastive in character: to know a claim P is to form a belief in P on the basis of a cognitive mechanism that enables the agent to discriminate the correct state of the world from a set of relevant not-P possibilities. Consider the following pair of examples.

> *Normal Barn Country.* While driving through the countryside, Henry points out various objects to his son: "That's a cow! That's a tractor! That's a silo! That's a barn!" Henry has excellent eye sight, and the objects he identifies are fully in sight and have the features characteristic of objects of the relevant kind. So, he identifies them correctly and confidently. Does Henry know that the last object he identified is a barn? The overwhelming temptation is to say, "Yes."
>
> *Fake Barn Country.* While driving through the countryside, Henry points out various objects to his son: "That's a cow! That's a tractor! That's a silo! That's a barn!" Henry has excellent eye sight, and the objects he identifies are fully in sight and have the features characteristic of objects of the relevant kind. So, he identifies them correctly and confidently. However, Henry is driving through a region used by Hollywood filmmakers to shoot scenes in a rural setting. The region therefore is populated by hundreds of fake barns which are perceptually indistinguishable from real barns. Unaware of this, Henry has

> unsuspectingly formed the true belief that the last object he points out is a barn. Does Henry know that the last object he identified is a barn? The overwhelming temptation is to say, "No." [33]

What explains the difference between the two verdicts? In *Normal Barn Country*, assuming that there are no fake barns around, if Henry forms a belief that the thing before him is a barn, on encountering a barn, we would indeed ascribe knowledge to him; for the perceptual capacity that he exercises in that scenario does put him in a position to distinguish the actual state of the world from a possibility where the content of his belief isn't true. However, in a scenario like *Fake Barn Country* where he is surrounded by fake barns, we wouldn't. For, in this scenario, there is a relevant alternative from which Henry is incapable of discriminating the correct state of the world: namely, the possibility that he is looking at a fake barn.

This lends support to the following conception of knowledge: in order to know a certain proposition, the agent must be able to distinguish the actual state of the world from certain relevant alternative possibilities where that proposition is false. Call this the *relevant alternatives theory* of knowledge. In a classic paper, Fred Dretske [1970] notes this feature of knowledge.

> To know that x is A is to know that x is A within a framework of relevant alternatives, B, C, and D. This set of contrasts, together with the fact that x is A, serve to define what it is that is known when one knows that x is A. (p. 1022)

He motivates this feature of knowledge by comparing it with the contrastive character of explanation.

> When I explain why Brenda did not order any dessert by saying that she was full (was on a diet, did not like anything on the dessert menu), I explain why she did not order any dessert *rather than, as opposed to,* or *instead of* ordering some dessert and eating it. It is this competing

[33]This example, originally credited to Carl Ginet, was first discussed in print by Goldman [1976].

> possibility which helps to define what it is that I am explaining when I explain why Brenda did not order any dessert...We explain why *P*, but we do so within a framework of competing alternatives *A*, *B*, and *C*...So it is with our epistemic operators. (pp. 1021-1022)

Later writers, like Alvin Goldman [1976], Gail Stine [1976], and David Lewis [1996] amongst others, have picked up on this theme. What they have ignored, however, is the analogy between explanation and knowledge that Dretske seems to be using here. Though Dretske perhaps doesn't think that the analogy goes any further than the structural similarity between knowledge ascriptions and statements of explanation, I think the contrastive character of knowledge is parasitic on the contrastive character of explanation.[34]

How? Let's focus on the TRUTH REQUIREMENT. According to the TRUTH REQUIREMENT, an agent's belief is free from veritic luck if and only if the basis of the belief explains why the agent holds a true belief rather than a false one. Now, whether or not the basis of a belief can satisfy this requirement will depend on the causal background against which the agent's cognitive processes operate.

In the *Fake Barn Country* example, the relevant causal background includes the fact that there are fake barns nearby. How should we individuate the basis of the agent's belief in such circumstances? Surely, in those circumstances, Henry could undergo the same perceptual experience as of there being an object with the characteristic properties of a barn before him, even if there were no real barn before him. So, at least on a picture that takes counterfactual dependence to be necessary (though not sufficient) for causal dependence, the formation of my belief doesn't causally depend on the presence of a real barn.[35] Now, a factor on which the formation of a belief doesn't causally depend cannot be included within the basis of

[34]For a similar point albeit based on a different approach to knowledge, see Rieber [1998].

[35]There might be a sense in which the formation of Henry's belief may causally depend on the presence of a real barn. If we think that Henry's perceptual experience has a demonstrative ingredient which picks out the barn before him, then that experience would indeed be object-dependent and therefore couldn't obtain unless there were a barn before him. Even if this is correct, this causally relevant fact need not necessarily be included in the basis of the agent's belief. Pre-theoretically, it seems that Henry doesn't know that the object before him is a barn, because his belief is at risk of error. But this couldn't be the case if the presence of the barn were included in the basis of Henry's belief.

that belief. So, the fact that Henry is looking at a real barn need not be included in the basis of his belief. Therefore, there are some possibilities compatible with the background conditions, where Henry forms a belief on the same basis, but he is in fact looking at a fake barn. That is why the basis of Henry's belief doesn't explain why the actual scenario is one where Henry truly believes that the object before him is a real barn, rather than a scenario where he is looking at a fake barn but believes that it is a real one. So, Henry doesn't know in this case.

The kind of explanation at work here is itself *contrastive.* In order for a belief to satisfy the TRUTH REQUIREMENT, the basis of the belief must explain why, among all the possibilities compatible with the background conditions, the actual scenario is one where the agent's belief is true, *rather than false.* These possibilities of falsehood, relative to which the truth of the belief is to be explained, are the *relevant alternatives* from which the agent must discriminate the truth in order to know it *tout court.* Any knowledge ascription, therefore, is always made relative to a set of relevant alternatives. When the basis of the agent's belief cannot explain why the agent believes P while being in a P-possibility rather than being in one of the not-P possibilities, she cannot be ascribed knowledge. This is exactly what happens in the *Fake Barn Country* scenario. Since the basis of the agent's belief cannot explain why the real barn possibility rather than the fake barn possibility obtains, she can't know that she is looking at a real barn rather than a fake one. Under this account, therefore, the fact that knowledge requires the elimination of relevant alternatives can be explained with reference to the contrastivity of explanation itself.

2.7.2 Reliability and Stability

Robert Nozick [1981] recognized two aspects of knowledge: any belief that counts as knowledge is both reliable and stable.

Nozick thought that a reliability condition was required to block cases of correctness luck. His preferred reliability condition was *sensitivity*: a belief in P formed by method M is *sensitive* to the truth of P if and only if, in all the nearby cases where P is false and the agent uses M to arrive at a belief about whether (or not) P holds,

the agent doesn't believe P by M. The sensitivity condition is concerned with the truth-conduciveness of the method that underlies a belief; for it says that in the nearby worlds where the content of the belief is false, the method by which the agent forms her actual belief wouldn't give rise to a belief in that content. That is why the sensitivity condition is a reliability condition. In *Fake Sheep*, Nina could easily have falsely believed the same claim by the same method; so, her beliefs are not sensitive to the truth.

Nozick also thought that a stability condition was required to rule out cases of defeat luck. He cast the stability condition in terms of *adherence*: a belief *adheres* to the truth if and only if, in all the nearby cases where P holds and the agent uses M to arrive at a belief about whether (or not) P holds, the agent believes P by M. Nozick [1997], tells us why the adherence condition is supposed to capture the *stability* of knowledge.

> In the tracking account of knowledge..., the fourth [adherence] condition states that if p were true, the person would believe it. The belief that p is stable under small enough perturbations. In terms of the modelling of subjunctives by possible worlds, he continutes to believe p in those p-worlds, the p-band of worlds closest to the actual world....Thus the fourth tracking condition gives us (a portion of) the requisite stability and stickiness of knowledge. (p. 151)

In *Assassination* and *Tom and Buck*, Jill and I respectively could easily have lost our beliefs formed by the same method upon encountering misleading evidence even if the contents of our beliefs were true. Our beliefs, therefore, don't adhere to the truth.

Neither sensitivity nor adherence are defensible.[36] Despite these problems for sensitivity and adherence, it seems plausible to think that more defensible relia-

[36]There is a huge discussion on the demerits of both *sensitivity* and *adherence*. For criticisms of sensitivity, see Goldman [1983], Vogel [1987], DeRose [1995], Williamson [2000, Chapter 7] and Kripke [2011]. Prominent objections to adherence have been given by Sosa [2002, p. 274], Kripke [2011, p. 178], Luper [2012], and Setiya [2012, p. 91]. As Kripke points out, a crushing objection against both these conditions is that they lead to radical failures of *epistemic closure*, i.e., the plausible thesis that competent deduction from known premises extends knowledge.

bility and stability conditions on knowledge could be constructed. So, the core insight that underlies Nozick's theory of knowledge survives: namely, that knowledge requires both a reliability condition and a stability condition, each of which is needed to block a distinct variety of knowledge-destroying epistemic luck.

EPISTEMIC EXPLANATIONISM shows that there is some truth to Nozick's dual aspect theory of knowledge: in order to be free from correctness luck as well as defeat luck, knowledge requires both reliability and stability. To see this, consider the following natural-sounding requirement on contrastive explanation.

> EXPLANATORY SUFFICIENCY. If C explains why an outcome E rather than an outcome F obtains, then F couldn't easily have occurred when C holds.

In other words, if C explains why E rather than F occurs, then C is sufficient to rule out the occurrence of F in the nearby cases.

To see why this is plausible, consider *indeterministic explanations*, i.e., explanations of events which have a non-zero objective chance of not happening. Suppose a tritium atom underwent spontaneous radioactive decay in an indeterministic world. If we want to explain this event, we may only be able to cite the fact that there was a probability of about three-fourths that it would. Though this might be an adequate explanation, this still doesn't explain why the atom underwent decay *rather than remaining intact.* More generally, the point is that there are no *indeterministic contrastive explanations.*[37] I think EXPLANATORY SUFFICIENCY accounts for this datum about contrastive explanations. When an event has a non-negligible chance of not happening, then there is a nearby possibility where the event doesn't happen. If there is such a nearby possibility, then an event which is compatible with that nearby non-occurrence of that event cannot explain why the event occurred rather than not occurring at all. In this case, the tritium atom could easily not have decayed even though it had a chance of three-fourths of decaying. That is why the chance fact cannot explain why the tritium atom decayed instead of remaining intact.

[37]Many writers acknowledge this. See Railton [1981], Lewis [1979], and Hitchcock [1999].

According to the TRUTH REQUIREMENT, the basis of the relevant belief explains why the belief is true, rather than false. Suppose EXPLANATORY SUFFICIENCY is true. If the basis of a belief explains why the belief is true, rather than false, then, in all the nearby cases, the basis of the belief leads to a true belief, rather than a false one. This gives us SAFETY FROM ERROR.

Now consider the POSSESSION REQUIREMENT. This says that a belief counts as knowledge only if the basis of the belief explains why it is held, rather than lost. According to EXPLANATORY SUFFICIENCY, if the basis of a belief explains why an agent holds a belief, instead of failing to hold it, then in all nearby scenarios where the agent forms a true belief on the relevant basis, the agent cannot fail to hold on to the belief. Now, if the evidential support that the belief in fact has were defeated by some piece of misleading evidence that in fact holds in the actual case, the agent would indeed lose her belief as long as her cognitive faculties are functioning properly. This implies that a belief satisfies the POSSESSION REQUIREMENT only if it is safe from evidential defeat. So, the POSSESSION REQUIREMENT, when taken in conjunction with EXPLANATORY SUFFICIENCY, entails the condition that we may label SAFETY FROM DEFEAT.

> SAFETY FROM DEFEAT. A belief formed on a certain basis amounts to knowledge only if, in every sufficiently similar case where the agent forms a belief on the same basis, the relevant belief is not rationally undermined by misleading evidence that holds in the actual scenario.

This is a stability condition on knowledge insofar as it requires the evidential support for the agent's belief to be undefeated in all nearby cases.

Thus, Nozick was right in thinking that both reliability and stability are required to block correctness luck and defeat luck. However, he was mistaken in thinking that they are sufficient for doing so. Even though these reliability and stability conditions provide good heuristic tools for detecting cases of correctness luck and defeat luck, they don't help us identify all cases of correctness luck and defeat luck. First, reliability need not always make a belief immune from correctness luck. Correctness luck doesn't just consist in a modally fragile relationship between the mo-

tivating reasons for an action or the basis of a belief and the correctness of the relevant action or belief. In cases like *Fake Sheep Redux*, where Nina is in hardworking sheepdog country, her belief suffers from correctness luck. Yet, the basis of her belief robustly leads to true beliefs in all nearby cases. Similarly, stability need not always block defeat luck. There may well be cases where the basis of a belief fails to explain why the belief is held rather than lost, but the belief is not at risk of evidential defeat. Therefore, while explanationism does give us arguments for reliability and stability conditions on knowledge, these conditions may at best only be imperfect counterfactual tests for deciding whether an example is an instance of correctness luck or defeat luck.

2.8 Conclusion

In this chapter, I have explored how an analogy between knowledge and moral worth can shed light on the nature of knowledge. On the one hand, it motivates a move away from EPISTEMIC RELIABILISM, the view that a belief is free from knowledge-destroying moral luck if and only if the basis of the belief reliably leads to the formation of true beliefs. On the other hand, it motivates a new anti-luck condition on knowledge, which I have called EPISTEMIC EXPLANATIONISM. According to this condition, a belief is free from knowledge-destroying moral luck if and only if the basis of the belief explains both why the agent has a true belief rather than a false one, and why she possesses the belief rather than lacking it. This account not only explains our intuitive judgements about cases, but also illuminates certain structural features of knowledge.

Before I close, let me highlight why this analogy between knowledge and moral worth is significant. First, the analogy is useful as a *tool of discovery*. It helps us see what might be going wrong with EPISTEMIC RELIABILISM in cases like *Fake Sheep Redux*, *Assassination*, and *Tom and Buck*. These examples, by themselves, do not constitute decisive objections against EPISTEMIC RELIABILISM: the reliabilist, for example, might be able to devise error-theories that explain our intuitions about these examples. However, once we notice the structural similarity between these

examples and the examples that cause trouble for reliabilism in the practical domain, we are able to see why notions like *safety* doesn't really get at the anti-luck condition that is necessary for knowledge: on the one hand, a belief could be safe from error due to factors about the agent's predicament which have nothing to do with the manner in which she forms her belief, and on the other hand, a belief could be safe from error even if it is lucky to remain undefeated by misleading evidence. By disclosing these problems for reliabilism, the analogy clears room for a more satisfactory account of knowledge-destroying epistemic luck, namely EPISTEMIC EXPLANATIONISM.

More generally, the analogy also vindicates the project of finding anti-luck conditions on knowledge, by revealing its broader significance. In the last few decades of epistemology, following Gettier's [1963] paper, many writers have sought a condition that would effectively block knowledge-destroying epistemic luck. The analogy between knowledge and moral worth shows that the phenomenon that these writers were investigating wasn't an isolated phenomenon that appears only in the epistemic domain; the kind of epistemic luck that destroys knowledge has analogues in the practical domain. In fact, we may expect that the kind of anti-luck condition that blocks knowledge-destroying luck will also block worth-undermining moral luck. Thus, the project of finding conditions that block knowledge-destroying epistemic luck will have significant consequences not just for epistemology, but also for moral philosophy.

3

Safety from Defeat

In Chapter 2, I showed that a condition called SAFETY FROM DEFEAT falls out of the explanationist conception of knowledge. According to this condition, roughly, a belief amounts to knowledge only if a belief formed on the same basis couldn't easily have been rationally defeated by misleading evidence. This is a *stability condition* on knowledge, insofar as it requires knowledge to involve belief that remains stable under small perturbations. In this essay, I explore the explanatory power of this stability condition on knowledge.

The chapter is divided into four parts. First, I say, again, why we need a stability condition like SAFETY FROM ERROR over and above a reliability condition like SAFETY FROM ERROR (§1). I then lay out the features of SAFETY FROM DEFEAT more clearly (§2). Next, I show why SAFETY FROM DEFEAT fares better than other stability conditions (§3). Finally, I show how SAFETY FROM DEFEAT explains a range of different epistemic phenomena. In particular, SAFETY FROM DEFEAT accounts for the explanatory role of knowledge in relation to certain kinds of behaviour, like rational perseverance (§4). It obviates certain demanding "internalist" conditions on knowledge (§5). It also illuminates the connection between

knowledge and practical interests (§6).

3.1 The Need for Stability

Recall Harman's [1973, pp. 142-144] examples.

> *Assassination.* A political leader is assassinated. An enterprising and reliable journalist witnesses the assassination and writes a report on it, which is then printed in the final edition of a newspaper. Jill reads the newspaper report, and comes to believe that the political leader is dead. But now the associates of the political leader, fearing a coup, decide to pretend that the bullet hit someone else. On nationwide television they announce that an assassination attempt has failed to kill the leader but has killed a secret service man by mistake. Rationally trusting this announcement, most people come to abandon their earlier beliefs about the death of the political leader. Luckily, Jill wasn't watching television when this announcement was broadcast. She continues to believe that the leader was assassinated.
>
> *Tom and Buck.* I am the library detective. One day, I see Tom smuggling a book out of the university library under his coat, and thus come to know this. I then testify before the University Judicial Council, saying that I saw Tom stealing a book from the library. After testifying, I leave the hearing room. Later that day, Tom's mother testifies at the same hearing. She claims that Tom couldn't have committed the theft; he was thousands of miles away at the time of the theft. However, she says, Tom's identical twin, Buck, was around, and he is an inveterate kleptomaniac.

For Harman, in each of these cases, a piece of evidence that the agent doesn't possess makes the agent lose her knowledge. In *Assassination*, surrounded by people who have evidence contrary to what she believes, Jill fails to retain her knowledge that the political leader is dead. In *Tom and Buck* everyone at the hearing rationally

believes that Tom didn't steal the book after Tom's mother testifies. The wide availability of information about Tom's mother's testimony destroys my knowledge that Tom stole the book.

As I pointed out, the beliefs in question don't suffer from *correctness luck*, the kind of luck makes a belief lucky to be true rather than false. In *Assassination*, the newspaper report that forms the basis of Jill's belief was produced by a reliable eyewitness. Similarly, in *Tom and Buck*, provided that Tom's mother is lying and Tom has no twin, my capacity to recognize people couldn't have failed me. So, it doesn't seem like a matter of luck that I hold a true belief rather than a false one.

This is also why a reliability condition, like SAFETY FROM ERROR, is of no help here.

> SAFETY FROM ERROR. A belief formed on a certain basis counts as knowledge only if, in every sufficiently similar case where the agent forms a belief on the same basis, the relevant belief is true.

In *Assassination*, the newspaper report that forms the basis of Jill's belief may be so utterly reliable that there are no nearby cases where it could have misled Jill about the political leader's death. Similarly, in *Tom and Buck*, provided that Tom's mother is lying and Tom has no twin, my capacity to recognize people might be working so well that there are no nearby cases where I mistake Tom for someone else.[1]

According to the diagnosis I offered earlier, in these scenarios, it is a matter of luck that the agent's belief isn't rationally defeated by misleading evidence.[2] That is why the agent seems to be in an epistemically precarious state. This kind of luck,

[1]In relation to *Assassination*, for example, Pritchard [2005, Chapter 6] seems to suggest that our hesitation to ascribe knowledge to Jill in this case can be explained by the fact that we don't in fact take Jill's belief to be free from risk of error. Since she lives in a state where the state interferes with the media, the belief that Jill forms on the basis of the newspaper report isn't reliable. It is not obvious to me that the state interference is an essential part of the description of the case. We could imagine a similar scenario where the public denial of the leader's death was a one-time prank that an otherwise globally reliable news channel chose to play on its viewers.

[2]As I understand *rational defeat*, the evidential support for a belief in a proposition P is *rationally defeated* at time t if and only if the agent's total evidence changes from E to E' at time t due to the rational impact of new evidence, such that the evidential support that E provides to P is not provided by E'. Under this construal, the evidential support for a belief cannot be rationally defeated by a fact, or a true proposition, unless the agent acquires that fact as new evidence. Such evidence

which is different from veritic luck, is what I shall call *defeat luck*.[3] To rule out instances of defeat luck, we need SAFETY FROM DEFEAT.

> SAFETY FROM DEFEAT. A belief formed on a certain basis amounts to knowledge only if, in every sufficiently similar case where the agent forms a belief on the same basis, the relevant belief is not rationally defeated by misleading evidence that holds in the actual scenario.

In *Assassination*, for example, Jill's belief that the political leader is dead could easily have been defeated by misleading evidence that is true in the actual scenario, i.e., by evidence about the television announcement. Similarly, in *Tom and Buck*, my belief could easily have been defeated by misleading evidence that is true in the actual scenario, i.e., by evidence about Tom's mother's testimony. In both these scenarios, therefore, my belief isn't safe from defeat.

The contrast between SAFETY FROM ERROR and SAFETY FROM DEFEAT lies in this. SAFETY FROM ERROR is what we may call a *reliability condition* on knowledge.

---

may either be *opposing evidence*, i.e., it may directly speak against the claim that the agent believes, or may be *undermining evidence*, i.e., it may work by attacking the connection between the evidence that supports the claim and the claim itself. For a classic discussion of rational defeat, see Pollock and Cruz [1999]. My description more closely follows Pryor [2013].

It is important to note how *rational* defeaters differ from other kinds of defeaters. First of all, rational defeaters are not what Bergmann [2006] calls *propositional defeaters*, because the latter kind of defeaters may not be part of the agent's evidence. In Harman's *Assassination* and *Tom and Buck* examples, the misleading evidence defeats the epistemic status of the belief, even though it is not acquired by the agent. That is why it is a propositional defeater, not a rational defeater. Second, rational defeaters are *normative defeaters*, as opposed to *doxastic defeaters*. A doxastic defeater against a belief in P is a claim Q that the agent in fact believes to be true, and which indicates that the agent's belief in P is false or unreliably formed. A normative defeater is a proposition Q that the agent ought to believe to be true, and which indicates that the agent's belief in P is false or unreliably formed. At least, if we assume that there is a tight connection between *having evidence* for a claim and *having reason* to believe it, rational defeaters are claims that the agent has evidence for, and therefore ought to believe to be true. In that sense, they count as normative defeaters.

[3] What I am calling *defeat luck* need not be completely distinct from evidential luck, but may just be a variety of evidential luck. A belief is evidentially lucky if and only if the agent is lucky to have the evidence she actually does. If an agent is lucky not to have encountered defeating evidence against a belief of hers, then her belief may be subject to evidential luck of some sort. Even if defeat luck is a form of evidential luck, this still doesn't mean that other forms of evidential luck—especially the kind of evidential luck that is instantiated in the cases described by Nozick and Unger—are incompatible with knowledge, or that defeat luck is compatible with knowledge.

Like other reliability conditions, it is concerned with the truth-conduciveness of the relevant belief-forming mechanism: it requires any belief that amounts to knowledge to be formed on a basis that doesn't lead to a false belief in nearby cases.[4] By contrast, SAFETY FROM DEFEAT is not concerned with truth-conduciveness, but rather with the robustness of the evidential support enjoyed by the content of the relevant belief. In this sense, it is what we may call a *stability condition* on knowledge: it requires any belief that counts as knowledge to be based on evidence that could not be rationally defeated in nearby cases by any fact about the agent's predicament.

3.2 Features of Safety from Defeat

Let me make the features of SAFETY FROM DEFEAT clearer by comparing it to SAFETY FROM ERROR.

3.2.1 Circularity

After Gettier [1963] offered his counterexamples to the traditional analysis of knowledge as justified true belief, epistemologists found themselves searching for analyses that would be immune to the Gettier-type counterexamples. The aim of this project was to find analyses of knowledge which would lay down *necessary and sufficient* conditions on knowledge that could be stated without *circularity*, i.e., in terms independent of our concept of *knowledge*. This project seems methodologically suspect in hindsight.[5] A better project might be to find necessary conditions on knowledge, which we may or may not be able to formulate in *knowledge*-free

[4]Williamson [2000, p. 124] clearly treats SAFETY FROM ERROR as a reliability condition. Sosa [1999] also takes it to be closely allied to Nozick's [1981] *sensitivity condition* which was intended to be a reliability condition on knowledge. However, it is important to note that SAFETY FROM ERROR can capture only one kind of reliability, namely *local reliability*, which involves error avoidance in scenarios sufficiently similar to the actual scenario. By contrast, *global reliability* relates to error avoidance in a range of different scenarios which may not be linked in any way to the actual scenario. For the distinction, see Goldman [1986].

[5]For criticism of this project, see Williamson [2000].

terms. Both SAFETY FROM ERROR and SAFETY FROM DEFEAT are *circular* necessary conditions on knowledge. Let me explain.

Both SAFETY FROM ERROR and SAFETY FROM DEFEAT can be subjected to a particular modal interpretation. A belief to be safe from error if and only if, in every sufficiently similar case where S forms a belief in P (or a sufficiently similar proposition P^*) on the same basis, the belief is true. Similarly, a belief is safe from defeat if and only if, in every sufficiently similar case where S forms a belief in P (or a sufficiently similar proposition P^*) on the same basis, the evidential support for the content of her belief is not rationally defeated by evidence about any fact that holds in α.[6] Thus, both of them appeal to a notion of *similarity* between *cases*.

In any particular case, whether a belief is reliable, or stable, will depend on what happens in sufficiently similar cases where the agent forms a belief on the same basis. Now, as Williamson [2009] points out in relation to SAFETY FROM ERROR, it would be a mistake to expect that we can spell out, independently of our pre-theoretic judgements about knowledge, what cases are sufficiently similar or what beliefs are formed on sufficiently similar bases. That is why both these conditions are circular necessary conditions on knowledge. This has an important consequence. It means that we cannot expect SAFETY FROM ERROR or SAFETY FROM DEFEAT to generate a clear prediction in every case, independently of whether we take the case to be an instance of knowledge.

To see this point in relation to safety from defeat, consider a variant of *Assassination*. In this case, too, Jill comes to believe that the political leader has died on the basis of the newspaper report, and a television announcement denying the news of his death has been broadcast on nation-wide television. However, a meteor is about to hit the earth. So, Jill won't have an opportunity to access the evidence about the television broadcast. Similarly, in all the nearby cases where Jill forms a belief on a sufficiently similar basis, but is prevented from accessing any misleading evidence because of the meteor, Jill's belief will remain rationally undefeated.

[6]Since α and α^* are centred possible worlds with a time-element in them, I have left the time-indices implicit in this modal interpretation of SAFETY FROM DEFEAT instead of explicitly mentioning them as I have done in the original statement of this condition.

However, in order to decide whether or not Jill's belief is in fact stable, we have to settle a more difficult question. Are there other nearby cases where Jill forms a belief on the same basis, but the belief is rationally defeated by evidence about her actual predicament?

This question cannot be settled without deciding whether her belief is stable enough to count as knowledge. Intuitively, it seems that it isn't: there is no epistemically significant difference between this case and Assassination. So, we should treat Jill's belief to be as unstable in this scenario as it is in Assassination. Hence, we have to acknowledge that there are nearby or sufficiently similar cases where there is no meteor that hits the earth, and Jill does get evidence that overturns her belief. In this sense, judgements about similarity and closeness will be parasitic on our pre-theoretic judgements about knowledge.

3.2.2 Modal Stability

SAFETY FROM ERROR is a *modal* reliability condition; it requires any belief that counts as knowledge to be formed on a basis that yields true beliefs across *nearby possibilities*. In this respect, it is different from a *diachronic* reliability condition which requires any belief that counts as knowledge to be formed on a basis that yields true beliefs across *different times*. Similarly, SAFETY FROM DEFEAT lays down a *modal* stability condition on knowledge, not a *diachronic* one. In other words, it doesn't say that when a belief amounts to knowledge, there won't be any *future time* at which a piece of evidence about the agent's predicament will rationally defeat the evidential support for the content of the belief. Rather, it says that there aren't any *nearby possibilities* where a belief, formed on a sufficiently similar basis, is rationally defeated by evidence about the agent's actual predicament.

To see the significance of this, consider an example from Williamson [2000], where the evidential support for a claim that an agent knows is rationally defeated by new evidence. Suppose at time t_1 I see a red ball and a black ball put into an empty bag. Thus, I come to know at that time that there are two balls in the bag, one red and the other black. I then see that in the first thousand draws with re-

placement, a red ball is drawn every time. This evidence, though compatible with what I earlier knew, should give me strong reason to doubt that there is a black ball in the bag. After these draws have been made, say at t_2, my total evidence will no longer support the claim that, of the two balls in the bag, one is black. The evidential support for what I knew will be rationally defeated.

Now consider the view that says that, for any belief that currently amounts to knowledge, there won't be any *future time* at which a piece of evidence about the agent's predicament will rationally defeat the evidential support for the content of the belief. Such a view would predict that my belief at t_1 wasn't knowledge. But this seems too strong: there is nothing intuitively wrong with my epistemic predicament at t_1. The right thing to say is that at t_1 I did know that two balls - one black and one red - were put into the bag, but then I rationally lost this knowledge. So, knowledge can be subject to future rational defeat.[7]

SAFETY FROM DEFEAT allows us to say this. In this case, there is a fact about my predicament that could rationally defeat the evidential support for my belief: namely, the fact that in the first thousand draws only a red ball will be drawn from the bag. But this fact isn't easily accessible to me before I witness the draws. There is no way I could have known in advance what the outcomes of the draws would be. That is why there are no nearby cases where my belief is rationally defeated by the relevant piece of evidence before I witness the draws. My belief is safe from rational defeat before the draws are made. In this manner, SAFETY FROM DEFEAT can allow a belief that counts as knowledge at one time to be rationally defeated at a later time.

[7]There is a substantive question as to how such rational defeat can occur. If we accept the view that whatever counts as knowledge is evidence, following Williamson's [2000] *E=K* thesis, then we have to make room for rational loss of evidence in order to make room for rational defeat of this kind. However, formal theories of belief-revision like Bayesianism do not seem hospitable to this phenomenon. For more discussion on this, see Weisberg [2009, 2015], Gallow [2014] and Greco [forthcoming]. As I say in Chapter 1, my account of higher-order defeat may be able to handle at least some such cases of rational defeat.

3.2.3 Diachronic Profile

Since SAFETY FROM ERROR picks out a reliability condition concerned with the truth-conduciveness of the cognitive mechanism that underwrites the relevant belief, the safety of a belief from error depends largely on its basis. However, SAFETY FROM DEFEAT lays down a stability condition that depends not only on the basis of the belief, but also on features of the agent's external environment which might not be included in the basis. This difference comes out quite clearly when we compare the diachronic profiles of the two conditions.

Suppose at t_1 an agent forms a belief which is safe from error. If at a later time t_2 she is able to retain her previous belief on the same basis as before, nothing epistemically significant about the causal history of the belief will have changed. If she was reliably connected to the world before, she will continue to be so even later. Hence, the nearby cases where the agent forms a belief on a similar basis, the belief cannot be false. So, the agent's belief will continue to be safe from error.[8]

By contrast, a belief that was previously safe from rational defeat could later come to involve a substantial risk of rational defeat, when the features of the agent's predicament change, even if she retains her belief on the same basis as before. To see why, recall *Tom and Buck*. In that scenario, before Tom's mother testified at the hearing, I did know that Tom stole the book. After she testifies, I don't, even though I hold a belief on the same basis as before. Here, my circumstances change, so that a piece of misleading evidence that wasn't available earlier becomes available, thus increasing the risk of rational defeat. So, the risk of rational defeat for a belief might vary across different times, even when the basis of the belief is held fixed. That is why it is important to index the risk of rational defeat for a belief

[8]Maria Lasonen-Aarnio [2010] defends this claim about the diachronic profile of SAFETY FROM ERROR. She uses it to support the view that, if safety from error is necessary and sufficient for knowledge, then past knowledge cannot be defeated when the agent retains her belief on the same basis as before. This, in turn, leads Lasonen-Aarnio [2013] to say that it is rationally permissible for an agent to disregard putative future counterevidence against a claim, if the agent currently knows that claim and is able to retain her belief on the same basis as before in the future. As a result, she embraces the conclusion of Kripke [2011]'s dogmatism puzzle, namely that, if an agent knows P, it is permissible for her to disregard any putative future counterevidence against P. I don't think this conclusion will hold in general if we admit other necessary conditions on knowledge.

to the agent's actual predicament, which is to be understood as a centred possible world with a time element in it.

3.3 Other Stability Conditions

Why is SAFETY FROM DEFEAT a good way to capture the stability condition on knowledge? To answer this question, I shall compare SAFETY FROM DEFEAT to two other principles which attempt to capture a similar condition: namely, *adherence* and *indefeasibility*.

3.3.1 Adherence

While offering his tracking account of knowledge, Robert Nozick [1981, p. 179] seems to acknowledge that knowledge requires both *reliability* and *stability*. Under his account, the reliability condition is captured by *sensitivity*: a belief in P formed by method M is *sensitive* to the truth of P if and only if, in all the nearby cases where P is false and the agent uses M to arrive at a belief about whether (or not) P holds, the agent doesn't believe P by M. By contrast, the stability condition is cast in terms of *adherence*: a belief *adheres* to the truth if and only if, in all the nearby cases where P holds and the agent uses M to arrive at a belief about whether (or not) P holds, the agent believes P by M. Nozick [1997] articulates the rationale behind the *adherence* condition in the following manner.

> In the tracking account of knowledge..., the fourth [adherence] condition states that if p were true, the person would believe it. The belief that p is stable under small enough perturbations. In terms of the modelling of subjunctives by possible worlds, he continutes to believe p in those p-worlds, the p-band of worlds closest to the actual world....Thus the fourth tracking condition gives us (a portion of) the requisite stability and stickiness of knowledge. (p. 151)

Now the adherence condition lives up to the rationale admirably in relation to *Assassination* and *Tom and Buck*. Since, in these cases, there is easily accessible misleading evidence that the agent could come to discover, there are a lot of sufficiently

similar cases where the content of her actual belief is true and she arrives at a belief with the same content using the same method, but she fails to hold on to the relevant belief. As a result, her belief fails to adhere to the truth.

Many have pointed out that the adherence condition cannot be right.[9] But there is a special reason for not being happy with it as a stability condition on knowledge. Let me explain with an example.

> *Proof.* I have proved a mathematical result, and come to believe the result on the basis of my proof. On the day I publish my proof, I happen to meet a colleague of mine - a mathematician whose opinion I respect a lot - who comes up to me, and says, "Congratulations on your proof!" However, unbeknownst to me, this mathematician's drink has been contaminated that morning, with a drug which, in most cases, makes people no better than chance at assessing mathematical proofs. Luckily, for me, the judgement of my colleague wasn't affected in the actual case, but there are many nearby cases where it was. In those cases, he wouldn't have congratulated me, and would instead have told me that the proof was wrong.

In *Proof*, it seems that I continue to know the mathematical claim even after I have published my proof. But there are nearby cases where the evidential support for my belief is defeated by the evidence that my colleague tells me that the proof is wrong. So, Nozick's adherence condition will predict that I don't know in this case. This seems like the wrong result.

In comparison, SAFETY FROM DEFEAT does better. A significant feature of SAFETY FROM DEFEAT is that it ties the counterfactual threats to knowledge tightly to facts about the agent's actual predicament: the sort of evidence that can threaten the possibility of knowledge must be true in the agent's actual epistemic predicament, and not merely in the counterfactual scenarios where it rationally defeats the agent's belief. In *Proof*, even though there are some nearby cases where the evidential sup-

[9]Prominent objections to the adherence condition have been given by Sosa [2002, p. 274], Kripke [2011, p. 178], Luper [2012], and Setiya [2012, p. 91].

port for my belief is rational defeated, the relevant piece of evidence - namely the evidence that my colleague tells me that the proof is wrong - isn't true in the actual predicament. The nearby possibility of such rational defeat doesn't threaten the epistemic status of the belief that I form in the actual case.[10]

3.3.2 INDEFEASIBILITY

Defenders of the defeasibility approach to knowledge also seem to impose a stability condition on knowledge: for them, knowledge requires indefeasible belief.[11] What does such *indefeasibility* consist in? On a simple defeasibility account, a belief amount to knowledge at time t only if there is no defeater for that belief at *t*, i.e., no fact at time *t* which, if discovered by the agent, would rationally defeat the evidential support for the content of her belief. Prima facie, this seems true about instances of possession luck, like Assassination and Tom and Buck, where a piece of evidence that the agent does not possess destroys her knowledge.

However, this is too strong. An obvious counterexample to this suggestion is provided by a variant of *Tom and Buck*, discussed by Harman (1973, p. 146), where I see Tom stealing the book and Tom's mother testifies in the court denying Tom's guilt, but everyone present knows that she is a pathological liar. In this scenario, there is available evidence that could rationally undermine my belief, but I continue to know.

To care of such cases, some writers, like Peter Klein [1981] state the indefeasibility condition as follows: a belief amounts to knowledge at time *t* only if there is no *ultimately undefeated defeater* for that belief at *t*. A defeater *E* is ultimately undefeated just in case there is no fact F such that the conjunction E and F justifies belief in *P*; or if there is such a fact, then there is some further fact F^* such that the

[10] In response, one might be tempted to restate *adherence* in the following form: a belief in *P* adheres to the truth in a case *a* just in case, in every nearby *P*-possibility where the agent uses method *M* to arrive at a belief about whether or not *P* holds, the agent doesn't fail to retain her belief in *P* in virtue of getting misleading evidence against *P* that is true in *a*. Note now this condition isn't significantly different from SAFETY FROM DEFEAT. This only shows that the truth that Nozick's adherence condition grasps at is best captured by SAFETY FROM DEFEAT.

[11] This approach has been defended by Lehrer [1965, 1969, 1974], Klein [1971, 1981], Swain [1974].

conjunction of E, F, and F^* fails to justify belief in P.[12] Since in *Tom and Buck*, there is a fact—namely, that Tom's mother is a pathological liar—which neutralizes the defeating force of the fact that Tom's mother says that Tom was in the library, there is no ultimately undefeated defeater in this case. So, I can continue to know that Tom stole the book.

Some writers have pointed out that this 'no ultimately undefeated defeater' condition is too weak: it is unable to rule out standard Gettier cases as defenders of this condition intend it to do.[13] For example, in the *Fake Sheep* example, there is a defeater for my belief—the fact that the object I see is a rock—but this fact, when conjoined with the fact that there is a sheep in the field, still provides adequate evidential support for the content of my belief. So, the defeater isn't ultimately undefeated, and hence the condition is satisfied.

More recently, John N. Williams [2015] has pointed out that defeasibility theories suffer from a more basic problem: they make first-personal knowledge of empirical knowledge impossible. Here is a simple way of putting the problem. Both the 'no defeater' condition as well as the 'no ultimately undefeated defeater' condition are universally quantified claims about facts: for example, the 'no ultimately undefeated defeater' condition says that a belief can satisfy these conditions only if every fact is such that it cannot rationally defeat the evidential support for the belief, or, if it does, it is ultimately defeated. Now, in the case of an empirical belief, it would be extremely difficult to ascertain whether the agent's belief satisfies such a condition. Note that *a posteriori* knowable facts can rationally impact the evidential support enjoyed by the contents of empirical beliefs. So, in order to determine whether an empirical belief satisfies the 'no defeater' condition or 'ultimately undefeated no defeater' condition, an agent must carry out some empirical investigation. However, it is unclear whether any amount of empirical investiga-

[12]Things are slightly more complicated; for there might be a further fact F^{**} such that the conjunction of E, F, F^*, and F^{**} justifies belief in P. To take care of such a fact, we would have to add a clause which says that there is a further fact F^{***}, such that the conjunction of all these facts fails to justify belief in P. And so on, *ad infinitum*. So, our characterization of an ultimately undefeated defeater is incomplete. However, this won't matter for the purposes of our discussion.

[13]See, for example, Turri [2012] and Foley [2012].

tion can tell whether all facts satisfy a certain condition. Hence, an agent won't be able to know, even by empirical investigation, whether an empirical belief of hers amounts to knowledge.

SAFETY FROM DEFEAT doesn't fall prey to these objections. First, unlike the 'no defeater condition', it is able to accommodate cases like the variant of *Tom and Buck*. Even in a scenario where potentially disconfirming evidence is easily accessible, the risk of rational defeat for a belief may be absent due to the presence of another easily accessible evidence which neutralizes the defeating force of the first piece of evidence. In the variant of *Tom and Buck*, there are some sufficiently similar possibilities where I come to know about Tom's mother's testimony. But everyone also knows Tom's mother to be a liar. So, the sufficiently similar possibilities in which I come to know about Tom's mother's testimony are also cases where I am given evidence about her habitual insincerity. As a result, the disconfirming evidence about Tom's mother's testimony won't rationally undermine my belief that Tom stole the book. That is why it can count as knowledge.[14]

Second, unlike the 'no ultimately undefeated defeater' condition, SAFETY FROM DEFEAT is able to rule out at least some Gettier cases. In the stopped clock case, I could easily discover that the clock has stopped working, and thus lose my belief due to the impact of that evidence. So, my belief isn't safe from defeat and therefore isn't knowledge. However, SAFETY FROM DEFEAT need not rule out all Gettier

[14]In relation to this variant of *Tom and Buck*, it is important to resist the temptation of accepting that there are nearby cases where information about Tom's mother's testimony rationally defeats my belief. This would mean that knowledge need not involve belief which is completely free from risk of rational defeat, but only one which is subject to a sufficiently small risk of rational defeat. According to this proposal, therefore, knowledge tolerates small risks of rational defeat. The problem with this proposal is that it makes room for failures of multi-premise closure, i.e., the principle that if an agent forms a belief by competent deduction from multiple known premises, while retaining the knowledge of those premises throughout, then her belief counts as knowledge. Since this proposal doesn't require complete freedom from risk of rational defeat, it allows accumulation of small risks of rational defeat in the course of competent deduction from multiple known premises. So, even though the agent's belief in each premise may be sufficiently safe from rational defeat to count as knowledge, her belief in the conclusion may not be. Thus, an agent might not always be able to extend her knowledge by competent deduction from multiple known premises. As John Hawthorne [2004, chapter 1] points out, multi-premise closure is explanatorily quite powerful. This should count as a structural consideration against this proposal.

cases: it isn't difficult to imagine a variant of the stopped clock example where a surprisingly effective conspiracy prevents me from easily discovering that the clock has stopped working. But this isn't a problem. As I pointed out, it is intended to rule out only cases of possession luck, not cases of veritic luck. Now, the standard Gettier cases are cases of correctness luck. So, it is a mistake to expect that SAFETY FROM DEFEAT will be able to eliminate all Gettier cases. In this respect, SAFETY FROM DEFEAT is a much less ambitious condition than what the defeasibility theorist takes the 'no defeater' condition or the 'no ultimately undefeated defeater' condition to be. Hence, the charge that it is too weak to rule out all Gettier cases doesn't have any force.

Third, SAFETY FROM DEFEAT is also able to accommodate the fact that in cases of empirical knowledge, we are often able to know that we know. Note that SAFETY FROM DEFEAT isn't a claim about all facts, but rather a claim about all sufficiently similar possibilities: it says that if an agent knows, then in every nearby possibility where the agent forms a belief on the same basis, the belief remains rationally undefeated by misleading evidence that is true in the actual scenario. How can an agent like us come to know whether or not such a condition obtains? Much in the same way we ascertain whether or not our beliefs suffer from a risk of error. When we are trying to determine whether our beliefs suffer from a risk of error, we gather information about the manner in which we form our beliefs, and our environment. Once we have decided that the manner in which a belief is formed couldn't have led to a false belief in the relevant circumstances, we declare it safe from error. Similarly, when we are trying to determine whether our beliefs suffer from a risk of defeat, we gather information about the manner in which we form our beliefs, and our environment. Once we have decided that our environment doesn't make our belief easily vulnerable to rational undermining, we take that belief to be safe from defeat.

3.4 Why Knowledge Explains Rational Perseverance

Internalists about mental states sometimes defend the claim that mental states depend solely on the intrinsic properties of the relevant agent on the basis of two premises. First, mental states causally explain behaviour. Second, only states that depend solely on the intrinsic properties of the agent can causally explain the behaviour of an agent.[15] Since the state of knowing does not depend solely on the intrinsic properties of the agent, it follows from the second premise that knowledge cannot causally explain behaviour. In response to this, some have claimed that knowledge can causally explain certain kinds of behaviour, such as rational perserverance. SAFETY FROM DEFEAT helps us account for this explanatory role of knowledge.

3.4.1 The Datum

Imagine a football coach being interviewed after his team has won its first victory of the season after a series of humiliating defeats. Relieved, the coach says,"We practised very hard for the last few months. But I *knew* it would pay off!" Perhaps, the coach is lying; perhaps, he did not know that the months of practice were going to result in a victory. The fact remains, however, that, by invoking knowledge, he is trying to give a rationalizing explanation of why his team persisted in working hard, even though they could not win a single match.

Williamson [2000, p. 62] notes this aspect of knowledge. He asks us to consider a burglar who, at great risk to himself, spends all night rummaging through a house in search of a diamond. Suppose that the burglar's cognitive faculties were in good order throughout, and therefore that he responded to her evidence well during the course of his search. Knowing nothing else about the burglar's situation, we might wonder: What was it about the burglar, when he entered the house, that made him rationally persevere in his search for the diamond? Williamson points out that the burglar's rational perseverance isn't explained well with reference to

[15]The *locus classicus* of this claim is Fodor [1987].

the fact the burglar *believed truly* that there was a diamond in the house and he wanted it. After all, the burglar might have formed this true belief after being told by someone trustworthy that it was under the bed in the master bedroom, when in fact it was in a drawer in the study. If the burglar had entered the house in search of the diamond with a true belief based on such a false premise, he could easily have lost his belief upon discovering that the diamond wasn't under the bed. By contrast, if we say that the burglar *knew* that there was a diamond in the house, we are immediately able to rule out the possibility that he formed his belief about the whereabouts of the diamond on the basis of a premise whose falsity he could easily discover upon entering the house. So, it is more probable that the burglar would rationally persevere in his search for the diamond, conditional on his knowing that there was a diamond in the house, than it is conditional on his believing truly that very same claim. Therefore, in examples of this sort, the agent's behaviour is better explained when we appeal to her state of knowing rather than when we merely appeal to her true beliefs.[16]

3.4.2 A Failed Proposal

We might ask: What feature of knowledge is it that makes an appeal to knowledge such a good explanatory move in these cases? A tempting response is this: it is the fact that knowledge must always involve belief that is safe from error. According to the discussion above, an appeal to a true belief doesn't explain the burglar's perseverance well, because the burglar could have formed that true belief on the basis of a false premise, in a manner similar to Gettier cases like *Fake Sheep*, and thus could easily have lost her belief upon discovering the falsity of the relevant premise. Since beliefs formed on the basis of false premises in Gettier cases like *Fake Sheep* are not safe from error, such beliefs are eliminated from the scope of knowledge by SAFETY FROM ERROR. So, we might think that it is the SAFETY FROM ERROR condition on

[16]Some, like Magnus and Cohen [2003] and Molyneux [2007] have objected to this observation. Their strategy is to show that there could be true beliefs which fall short of knowledge, but could still play the same explanatory role that knowledge plays in these cases. For a rebuttal of these objections, see Nagel [2013].

knowledge that accounts for the explanatory role that knowledge plays in relation to rational perseverance.

This proposal cannot work. If the burglar's belief was safe from error when he entered the house, then in all nearby cases where he formed a belief on a similar basis, his belief would be true. But whether or not a belief is rationally defeated by misleading evidence has nothing to do with its truth! So, the safety of the burglar's belief from error at the time of his entering the house doesn't tell us why some piece of misleading evidence didn't make him lose his belief due to rational defeat. Even if the burglar's belief had not been formed on the basis of false premises and had been completely safe from error, the burglar could still have lost his belief pretty quickly after entering the house, had there been easily discoverable misleading evidence lying around in the house.[17]

3.4.3 A Better Solution

For the burglar to have rationally persevered in his search for the diamond throughout the night, the evidential support for the content of his belief must have remained rationally undefeated during the course of his search. So, knowledge can explain rational perseverance only if there is a feature of knowledge that makes it relatively (but perhaps not completely) invulnerable to rational defeat. I want to suggest that this feature is none other than the one captured by SAFETY FROM DEFEAT. According to this condition, any instance of knowledge always involves a true belief that is safe from rational defeat, a true belief that remains rationally undefeated by misleading evidence in all nearby cases. So, if the burglar knew at the time of entering the house that there was a diamond in the house, then his belief at that time couldn't easily have been defeated by misleading evidence. This could only have been the case if the burglar, at the time of entering the house, was in an environment which didn't contain any easily accessible misleading evidence. Now,

[17]Williamson [2000] himself seems to acknowledge this: "Variants of the previous case can be constructed in which the burglar enters the house believing truly that there is a diamond in it without reliance on false lemmas, yet fails to know in virtue of misleading evidence which he does not then possess, but may discover in the course of his search, in which case he will abandon the search" (p. 63).

if the burglar was in such an environment, it is quite likely that his belief would persist into the future, and therefore that he would rationally persevere in his search for the diamond. That is why an appeal to the burglar's knowledge is a better explanation of his rational perseverance than an appeal to his true belief. This is how SAFETY FROM DEFEAT accounts for the explanatory role that knowledge plays in relation to rational perseverance.[18]

3.5 KNOWLEDGE AND INTERNAL PERSPECTIVE

Some take reliability to be necessary and sufficient for knowledge. Call this view *simple reliabilism*. Against this view, others argue that the agent's internal perspective on the reliability of the belief-forming mechanism underlying a belief may prevent that belief from amounting to knowledge, even when it is reliably formed. This yields a certain "internalist" condition on knowledge. SAFETY FROM DEFEAT helps us see that even though simple reliabilism may not be correct, there is still no need to accept such "internalist" conditions on knowledge.

3.5.1 THE KNOWLEDGE-ACCESS PRINCIPLE

Consider Laurence Bonjour's [1985] case of Norman the clairvoyant:

> Norman, under certain conditions which usually obtain, is a completely reliable clairvoyant with respect to certain kinds of subject matter. He possesses no evidence or reasons of any kind for or against the general possibility of such a cognitive power or for or against the thesis that he possesses it. One day Norman comes to believe that

[18]This, however, is not to say that if the burglar knows that there is a diamond in the house at the time of entering the house, his belief *won't* be defeated by future evidence that he gains in the course of ransacking the house. Since SAFETY FROM DEFEAT isn't a diachronic stability condition, it leaves open the possibility that the burglar's knowledge about the whereabouts of the diamond might be rationally defeated at some point in the future. It only requires that there be no nearby cases where the burglar's belief is rationally defeated. Thus, for the burglar's belief to be safe from rational defeat at the time of entering the house, it only has to be the case that any evidence that might rationally defeat the burglar's belief in the future isn't something that he could easily access at that time.

> the President is in New York City, though he has no evidence either for or against this belief. In fact the belief is true and results from his clairvoyant power under circumstances in which it is completely reliable. (p. 41)

Here, a simple reliabilist who takes reliability to be both necessary and sufficient for knowledge will be forced to say that Norman does know that the President is in New York.[19] But, intuitively, it seems as if he doesn't.

Bonjour thinks that this is because Norman isn't propositionally justified in believing that the President is in New York.[20]

> From his standpoint, there is apparently no way in which he could know the President's whereabouts. Why then does he continue to maintain the belief that the President is in New York City? Why isn't the mere fact that there is no way, as far as he knows, for him to have obtained this information a sufficient reason for classifying this belief as an unfounded hunch and ceasing to accept it? And if Norman does not do this, isn't he thereby being epistemically irrational and irresponsible?...Part of one's epistemic duty is to reflect critically upon one's beliefs, and such critical reflection precludes believing things to which one has, to one's knowledge, no reliable means of epistemic access. (p. 42)

On the basis of this diagnosis, Declan Smithies [ms.] has recently claimed that this shows that a belief must be capable of surviving ideal critical reflection in order to count as propositionally justified. Ideal critical reflection is the kind of critical reflection that will be undertaken by an ideal counterpart of an agent who doesn't suffer from the computational or conceptual limitations that the agent is actually

[19]The original target of this objection was Armstrong's [1973] simple reliabilist account of non-inferential knowledge. Goldman [1986] defends himself against this objection by embracing a weaker version of reliabilism, under which reliability is necessary, but not sufficient for knowledge and justification.

[20]To have *propositional justification* for a belief is to have adequate epistemic reason for holding that belief. To be *doxastically justified* in holding a belief, the agent must not only have adequate epistemic reason for holding it, but must also base her belief properly on that reason. Here, I am roughly understanding *epistemic reason* in terms of considerations that are part of the agent's evidence.

subject to.[21] In this scenario, Norman has no way of reflectively vindicating the epistemic credentials of his belief, even if we were to beef up his computational or conceptual capacities infinitely while holding fixed his evidence.[22] So, his belief cannot withstand ideal critical scrutiny, and therefore cannot count as justified.

For Smithies, in order to be propositionally justified, an agent's belief must be capable of surviving an idealized process of critical reflection. This entails a form of *access internalism* about propositional justification, where in order to have propositional justification for a belief an agent must have reflective access to the justificatory status of the belief.[23] If we now take propositional justification to be a necessary condition on knowledge, then we get the following constraint.

> *Knowledge-Access Principle.* If a belief in P counts as knowledge, then the relevant agent must have propositional justification to believe upon ideal critical reflection that she has propositional justification to believe P.

This means that the epistemic status of a belief depends on the internal perspective of the agent on its epistemic credentials.

3.5.2 The Problem

Whether or not we accept access internalism about justification, it is not obvious to me that knowledge is subject to such an internalist justification condition. My

[21] Bonjour's original discussion suggests that, in order to be justified in holding a belief, an agent must be actually or potentially aware, on having critically reflected in a manner that is responsible, that her belief is justified. As subsequent discussion, e.g. Bergmann's [2006] discussion of what he calls the *subject's perspective objection*, has revealed, this is too strong. Smithies modifies this requirement to the weaker requirement that the agent must have *propositional justification* to believe that her belief is justified, where propositional justification doesn't require actual or potential awareness given the agent's *actual* cognitive capacities. Rather, it involves what the agent would believe after ideal critical reflection.

[22] It is not obvious whether this in fact is possible; for enriching an agent's conceptual repertoire will inevitably enhance her capacity for discriminating possibilities in modal space. That in turn might give her new modal information. But we can set this complication aside for now.

[23] For a similar claim about *doxastic justification*, see Alston [1988], and for elaboration of Smithies' own view, see Smithies [2012, forthcoming].

worry is this. It seems that the scenario that Norman is in isn't significantly different from one where an infant for the first time forms a perceptual belief about the external world.[24] If Norman doesn't have any reason to endorse his clairvoyant beliefs upon ideal critical reflection, then neither can the infant have justification upon ideal critical reflection to believe that her perceptual beliefs are justified. Let me explain.

The infant doesn't have any empirical evidence in favour of the reliability of her perceptual faculties. So, she simply doesn't have the requisite information to recognize what her belief has going for it.[25] Without evidence about the reliability of her perceptual faculties, the infant cannot conclude, even on ideal critical reflection, that she is justified in holding her perceptual belief. After all, if she doesn't have enough evidence to discount the possibility that her perceptual faculties are unreliable, she cannot justifiably take her perceptual experiences to give her sufficient evidence to believe claims about the external world.[26] Yet, provided that her belief is reliably formed, it would be hard to resist the temptation to ascribe

[24]To make the analogy strong, we could imagine that clairvoyance, as a belief-forming mechanism, works exactly like perception, by yielding quasi-perceptual awareness of events that are spatio-temporally distant from the relevant agent.

[25]This is true only if we assume that an agent cannot be justified, independently of any empirical investigation, to take her perceptual faculties to be reliable. Some writers, such as Crispin Wright [2004] and Roger White [2006], question this assumption. I have two comments in response to this. First, it is unclear to me whether we can have non-empirical justification for believing that our perceptual faculties are reliable. Second, even if we do have non-empirical justification for believing in the reliability of perception, why can't a similar kind of justification be available to Norman for believing that his powers of clairvoyance are reliable? The more general point is that, given the similarity between Norman and the infant, whatever we say about the one case will also apply to the other.

[26]This is true on both Cartesian and externalist accounts of evidence. On the Cartesian account, an agent's evidence consists in facts about or events pertaining to her phenomenal states. See Conee and Feldman [2004] for such a view. If an agent is to justifiably believe a phenomenal state to be evidence for a claim about the external world, she must have background evidence to think that the relevant phenomenal state favours the truth of the relevant claim. However, if she has no evidence to think a phenomenal state is reliable, then she doesn't have such background evidence. By contrast, on typical externalist accounts of evidence, like Williamson's [2000] *E=K* thesis or Goldman's [2009] reliabilist account, the reliability of a belief-forming mechanism is required for the deliverances of that belief-forming mechanism to count as evidence. So, if an agent doesn't have good reason to take a belief-forming mechanism to be reliable, she cannot take its deliverances to be evidence for any claim.

knowledge to her. How can we then account for the difference between Norman's epistemic predicament and that of this infant?

So, we face a dilemma between two extreme views. On the one hand, there is the simple reliabilist view, according to which reliability is necessary and sufficient for knowledge. This view forces us to say that Norman does know that the President is in New York. On the other hand, there is the internalist view, according to which an agent can only know if her belief is capable of surviving ideal critical reflection. This view predicts that an infant who doesn't know anything about the reliability of her perceptual faculties cannot thereby gain knowledge by means of those faculties. Since both views have unpalatable consequences, we should avoid both. But, at least on the face of it, these seem to be the only theoretical options here.

SAFETY FROM DEFEAT offers us a third option: it helps us explain why Norman doesn't know, but the infant does. The story is this.

3.5.3 The Social Dimension of Knowledge

If we take Norman to be a member of an epistemic community like ours, where information gathered by putative powers of clairvoyance is treated with suspicion, then Norman could easily have gained evidence that people who claim to have direct knowledge like him about spatially or temporally distant things are generally unreliable charlatans. This would give him reason to doubt the epistemic credentials of his clairvoyant beliefs. In that case, Norman would rationally lose the belief that he forms in the actual case. So, Norman's actual belief isn't safe from rational defeat, and therefore doesn't amount to knowledge. But this instability has nothing to do with whether Norman has justification on ideal critical reflection to believe that his belief about whereabouts of the President is justified. It has to do with the fact that we expect Norman to be a member of a community where there is misleading evidence in the air, which could easily rationally defeat his belief.

A similar response can be given to other examples of this kind, where absence of evidence for the reliability of a belief-forming mechanism supposedly prevents a belief formed by that mechanism from counting as knowledge. Take Lehrer's

[1990] example of Mr. Truetemp who has a temperature-detecting device implanted in his brain, which reliably generates true beliefs about the ambient temperature. Lehrer wants to deny that these beliefs count as knowledge. Once again, we have an explanation from SAFETY FROM DEFEAT, which spells out why this is so. Assuming that Mr. Truetemp belongs to a community like ours, he is likely to be exposed to misleading evidence against the epistemic credentials of his beliefs, if he were to tell others about how he came to know about the temperature without relying on the measurements of a thermometer or similar accredited ways of detecting temperature. Since his beliefs are thus vulnerable to rational defeat, they can't count as knowledge.

What distinguishes the case of the infant from these cases is that our community is also a community where perception is recognized as an epistemically respectable means of forming beliefs. Since we assume that the infant belongs to our community (as we do in the cases of Norman and Mr. Truetemp), we take her to be in an environment where her first perceptual beliefs aren't exposed to any misleading evidence that casts doubt on their epistemic credentials. So, even though they are formed in the absence of any empirical evidence for the reliability of her perceptual faculties, they are safe from rational defeat. That is why they can still count as knowledge.[27]

Thus, SAFETY FROM DEFEAT allows us to escape the consequences of two extreme views, namely simple reliabilism and the internalist conception of knowl-

[27]Two clarificatory points. First of all, this doesn't mean that knowledge is independent of higher-order evidence, i.e., evidence about the epistemic credentials of the relevant agent's beliefs. In fact, in my explanations of Norman the clairvoyant and Mr. Truetemp cases, I am assuming that the propositional justification for a belief can be defeated by misleading evidence against the reliability of the relevant belief-forming mechanism. If my assumption is correct and knowledge requires propositional justification, an agent's knowledge can indeed be undermined by misleading higher-order evidence. For the view that higher-order evidence can rationally defeat the justification for a belief, see Feldman [2005], Christensen [2010] and Horowitz [2014]. Some writers, such as Williamson [2011], Wedgwood [2012], and Lasonen-Aarnio [2014], are sceptical of this. Second, a further interesting question is where this leaves us with access internalism about justification. There are two options here. On the one hand, one could accept access internalism about justification and deny that knowledge requires justification. This might seem implausible to some. On the other hand, one could reject access internalism about justification, and hold on to the thesis that knowledge requires justification. I find the second option more acceptable than the first.

edge laid down by the *Knowledge-Access Principle*. It does this by appealing to the insight that, in scenarios like that of Norman the clairvoyant, it is not the internal perspective of the agent that prevents her from gaining knowledge. It is rather the perspective of her community that proves to be an impediment. In this respect, the account of knowledge that falls out of SAFETY FROM DEFEAT is comparable to an *anti-individualist* account of mental content defended by Tyler Burge [1979, 1986]. This account, like other forms of anti-individualism, acknowledges that the contents of an agent's thoughts may not be determined solely by what is in her head. But it differs from other anti-individualist accounts like that offered by Putnam [1975]; for it says that those contents are determined not just by her causal relationships with her physical environment, but also by her social environment, i.e., the practices of her linguistic community. Analogously, if SAFETY FROM DEFEAT is right, then what an agent knows is fixed not only by the causal connection between her belief and her physical environment, but also by the practices and beliefs of the community she belongs to.

3.6 PRAGMATIC ENCROACHMENT

It is tempting to think that whether an agent knows depends solely on factors that bear upon the truth of her belief, e.g., the evidence that she possesses, or the reliability of the cognitive mechanism that she uses to form the relevant belief. Call such factors *alethic* factors. Some, however, have argued that this isn't true; an agent's knowledge could depend on her *practical interests*, over and above alethic factors like evidence or reliability.[28] This is sometimes called *pragmatic encroachment* on knowledge. I want to show how SAFETY FROM DEFEAT can help us explain at least some cases that motivate pragmatic encroachment, while avoiding the theoretical costs incurred by standard defences of pragmatic encroachment.

[28]For a defence of this view, see Fantl and McGrath [2002, 2007, 2009], Hawthorne [2004], Stanley [2005], and Hawthorne and Stanley [2008].

3.6.1 Examples

Consider the following examples described by Stanley [2005].

> *Low Stakes.* Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. It is not important that they do so, as they have no impending bills. But as they drive past the bank, they notice that the lines inside are very long, as they often are on Friday afternoons. Realizing that it isn't very important that their paychecks are deposited right away, Hannah says, 'I know the bank will be open tomorrow, since I was there just two weeks ago on Saturday morning. So we can deposit our paychecks tomorrow morning.'
>
> *High Stakes.* Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since they have an impending bill coming due, and very little in their account, it is very important that they deposit their paychecks by Saturday. Hannah notes that she was at the bank two weeks before on a Saturday morning, and it was open. But, as Sarah points out, banks do change their hours. Hannah says, 'I guess you're right. I don't know that the bank will be open tomorrow.'
>
> *Ignorant High Stakes.* Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since they have an impending bill coming due, and very little in their account, it is very important that they deposit their paychecks by Saturday. But neither Hannah nor Sarah is aware of the impending bill, nor of the paucity of available funds. Looking at the lines, Hannah says to Sarah, 'I know the bank will be open tomorrow, since I was there just two weeks ago on Saturday morning. So we can deposit our paychecks tomorrow morning.' (pp. 3-5)[29]

[29] Cases of this kind were first discussed by DeRose [1992]. I have included *Ignorant High Stakes* in my discussion of pragmatic encroachment for a certain reason. Brian Weatherson [2005] has argued that a pragmatic account of belief can take care of certain putative cases of pragmatic encroachment like *High Stakes*. However, as Weatherson [2012] later acknowledges, this explanation doesn't work for a scenario like *Ignorant High Stakes* where the agent is unaware that she is in a high stakes scenario. In order to address the phenomenon of pragmatic encroachment fully, it is important to have cases like *Ignorant High Stakes* on the table.

The commonly reported intuition is that Hannah's self-ascription of knowledge is correct in *Low Stakes*, but not in *High Stakes* and *Ignorant High Stakes*. Why? Note that we cannot explain the difference by appeal to any alethic factor, like Hannah's evidence or the reliability of the underlying belief-forming mechanism: Hannah might be equally reliable and have the same evidence about the bank's office hours in each of the cases. SAFETY FROM ERROR therefore cannot be of any use here.

A standard strategy is to explain this difference by arguing that there are pragmatic constraints on propositional justification, i.e., on what counts as *sufficient* evidence for a belief.[30] Since the practical stakes are higher in *High Stakes* and *Ignorant High Stakes*, one might say, the evidential support that the content of Hannah's belief enjoys isn't strong enough for her belief to be propositionally justified in those cases, but is sufficient for propositional justification in *Low Stakes*. This diagnosis conflicts with *evidentialism* about propositional justification, i.e., the view that the propositional justification that an agent has for a belief depends just on the evidence that she possesses. But this seems like a bad consequence: why should a pragmatic factor that has nothing to do with the truth of the belief should affect its justificatory status? If we like evidentialism, we should try to account for the influence of practical interests on knowledge without relying on any controversial claim about propositional justification.

3.6.2 The Possibility of Pragmatic Encroachment

Here, SAFETY FROM DEFEAT can come to the rescue. If SAFETY FROM DEFEAT is correct, then factors like practical interests which have nothing to do with the truth of the relevant belief can determine the risk of rational defeat that a belief involves. That is why such factors can affect the epistemic status of a belief.

Recall Williamson's burglar case. Let's say that the burglar was not the only one to have formed the belief that there was a diamond in a particular house. There was another person, the burglar's *alethic twin*, who formed a belief in the same proposition on a sufficiently similar basis. If knowledge were solely a matter of alethic fac-

[30]For this view, more generally, see Fantl and McGrath [2002, 2009].

tors, then the following biconditional would hold: the burglar's *alethic twin* knows that there is a diamond in the house if and only if the burglar knows that there is a diamond in the house. But SAFETY FROM DEFEAT shows that this might not be true. Let us imagine that the practical interests of the burglar's alethic twin are very different from those of the burglar: the twin would never go seeking the diamond in the house. While there are nearby cases in which the burglar searches for the diamond in the house, there are no nearby cases where his alethic twin does so. Obviously, as we saw in Williamson's example, in engaging in such activities, the burglar exposes himself to certain kinds of misleading evidence, which might defeat the evidential support that his belief enjoys. By contrast, the burglar's alethic twin does not expose himself to such misleading evidence in nearby cases. We could now easily imagine a scenario where the belief formed by the burglar's alethic twin is safe from rational defeat, but the burglar's belief is not. In such a scenario, the burglar's alethic twin might know that there is a diamond in the house, but the burglar won't.

This shows that the practical interests of an agent can determine what kinds of evidence she will be exposed to in nearby cases. Since practical interests can in this manner regulate the risk of rational defeat that the agent's beliefs involve, they can prevent some of these beliefs from amounting to knowledge. Now, let us return to *High Stakes* and *Ignorant High Stakes*. If we think that *modal information* can rationally defeat the evidential support for the content of a belief, then we can show how Hannah's practical interests in *High Stakes* and *Ignorant High Stakes* increase the risk of rational defeat for her belief that the bank will be open the next day, by exposing her to certain kinds of modal information in nearby cases. This would tell us why Hannah doesn't know in these cases. Let me explain.

3.6.3 EXPLAINING THE EXAMPLES

Sometimes, new modal information can have defeating force against a previously rational doxastic attitude. For example, Einstein's discovery of the general theory of relativity in 1915 rationally undermined the confidence of the scientific com-

munity in the Newtonian theory of gravitation. Presumably, what brought about the shift of confidence wasn't just evidence about how things *are* in the world, but rather evidence about a previously undiscovered possibility, i.e., Einstein's theory. When scientists realized that this new theory explained the available evidence better than the Newtonian theory, they came to reject the latter theory.[31] Intuitively, at least, the confidence that the scientific community had in the Newtonian theory of gravitation was rational before Einstein's discovery of the general theory of relativity. So, the modal information made accessible by Einstein's discovery rationally defeated the evidential support for the content of a previously rational doxastic attitude.

The kind of possibility that the scientific community became aware of in this case ought to be distinguished from two other kinds of possibility which do not have such defeating force against previously rational doxastic attitudes. First of all, there are possibilities that an agent *ought to attend to* while forming beliefs about a certain subject-matter; if she were to form a belief about that subject-matter without attending to them, her belief would not even be rational.[32] Second, there are possibilities that are so *implausible* that discovering them cannot have any defeating force against the agent's beliefs.[33] When an agent becomes aware of possibilities of these two sorts, the newly gained modal information *cannot* rationally undermine a previously rational belief formed by disregarding them: if the ignored possibil-

[31]Some have argued that standard theories of belief-revision like Bayesianism cannot capture the rational impact of such modal information. This has been labelled the *problem of new theories* by John Earman [1992].

[32]Take an example. Suppose I am about to enter a room and a friend whom I rationally trust lies to me, saying, "The room is lit up with red light that makes any surface look red." As I enter the room, I undergo a perfectly reliable experience as of a red wall before me, but I also stub my toe on a nail. Distracted by the pain, I don't even consider the possibility that the wall that I see before me might be white with a red light shining on it. I just form the belief that the wall before me is red. Given the evidence I have, it seems that I ought to attend to this possibility in this scenario; failing to do so seems to make my belief irrational.

[33]For example, if I am considering whether I will make it to a meeting on time, it seems that I can make up my mind about this without paying any attention to the possibility that I will be kidnapped by aliens on my way to the meeting. So, even if I were to disregard this possibility and form the belief that I would make it to my meeting on time, then newly gained awareness of this outlandish possibility could not rationally defeat the evidential support for the claim that I believe.

ities were ones she ought to have attended to, her previous belief wouldn't have been rational in the first place, and if the ignored possibilities had been outlandish, the newly gained modal information would have no defeating force.

By contrast, accessing possibilities like Einstein's theory *can* rationally undermine a previously rational attitude. Why? First of all, possibilities of this kind need *not be readily accessible* to agents like us who only have limited computational resources.[34] So, it could indeed be rationally permissible for such agents to ignore possibilities of this kind. If a computationally limited agent adopts a doxastic attitude without attending to or reflecting on certain hypotheses that are rationally permissible for her to ignore in the relevant scenario, the doxastic attitude might indeed be regarded as rational in the light of her computational limitations. Second, once the agent comes to access possibilities of this kind, they may be *plausible* enough by her lights to be taken seriously. If it then turns out that these possibilities account for the agent's evidence as well as the hypothesis she formed her doxastic attitude about, then she might indeed lose her reason for holding her previous doxastic attitude. That is how modal information of this sort can rationally defeat the evidential support for the content of previously rational doxastic attitudes.

I want to claim that in *High Stakes* and *Ignorant High Stakes,* Hannah is in a predicament where she could easily come to discover possibilities of this third kind. In *High Stakes,* Hannah and Sarah know that they are in a scenario where failing to deposit the check in time will prevent them from paying the impending bill on time. Since Hannah cares about paying the bill, she should be more anxious

[34]One way for a possibility to fail to be readily accessible to an agent is for it to be conceptually indiscriminable by her: the general theory of relativity wasn't readily accessible to the scientific community before Einstein's discovery, because the scientific community lacked the repertoire of concepts required for formulating the theory. However, the kind of modal information that I am talking about here need not necessarily be information about possibilities that the agent could not conceptually discriminate before; the relevant possibilities might be just ones which the agent isn't capable of accessing due to other kinds of computational constraints, e.g., constraints of time. This kind of modal information has received discussion in two different kinds of literature. First, economists, e.g., Modica and Rustichini [1994] and Dekel, Lipman and Rustichini [1998], have discussed the absence of such modal information under the label *unawareness.* Second, in the literature on epistemic modals, Ciardelli *et al.* [2009] and Yalcin [2011] have tried to show that utterances of sentences involving 'might' impact the conversational context by calling attention to possibilities to be taken seriously in the context of conversation.

about the bank hours. Due to her anxiety, she could easily come to entertain certain sceptical possibilities where her evidence about the bank hours isn't reliable, e.g., the possibility where the bank changes its hours.[35] Given that Hannah is an agent like us who usually operates under computational constraints, she might not have considered these possibilities while forming her belief that the bank would be open the next day. Assuming that it was then rationally permissible for her to ignore these sceptical possibilities, we might think that she indeed was epistemically rational in forming her belief. However, once she comes to entertain these sceptical possibilities, these possibilities, e.g., the possibility where the bank changes its hours, might be plausible enough by her lights to be taken seriously. Since they also explain Hannah's evidence as well as the claim she believes, the evidential support for the content of her belief might be rationally defeated. Since Hannah is in a scenario where she could easily become aware of such possibilities due to her anxiety about the impending bill, her belief is exposed to a risk of rational defeat. That is why she doesn't know that the bank will be open the next day.

Similarly, in *Ignorant High Stakes*, even though Hannah and Sarah don't know about the impending bill, Hannah could easily discover this fact about her predicament. If she did, she would once more become aware of sceptical possibilities where her evidence about the bank hours isn't reliable. Here, too, Hannah's belief is subject to a risk of rational defeat. By contrast, in *Low Stakes*, there is no such impending bill, so there are no nearby cases where Hannah becomes aware of such sceptical possibilities and loses her belief. Hannah's belief, therefore, doesn't involve any risk of rational defeat. Thus, in *High Stakes* and *Ignorant High Stakes*, it is Hannah's practical interests that prevent her from knowing the relevant claim.

The success of this approach depends on the thesis that there is a kind of modal

[35]In relation to certain cases that cause trouble for principles of epistemic closure, John Hawthorne [2004, chapter 4] discusses the role of anxiety induced by high stakes in making certain sceptical possibilities salient. Along the same lines, Jennifer Nagel [2011] uses a distinction between intuitive and reflective modes of cognition to suggest that, when stakes are high, the reflective mode of cognition is brought into play, and this in turn makes certain sceptical possibilities more salient than they previously were. It is not clear whether any of these writers think that the salience of such possibilities has any defeating force against the belief that the agent held earlier. I am suggesting that this is in fact the case.

information that can rationally defeat the evidential support for the content of a belief. This idea, of course, needs to be defended in more detail elsewhere. But let me close by pointing out an distinctive feature of this approach: even though it allows practical interests of an agent to encroach on knowledge, such interests encroach on knowledge only by determining what kinds of misleading evidence (modal or non-modal) the agent might encounter in nearby cases. This has two important consequences. First, it implies that if an agent's evidence is robust enough not to be defeated by the misleading evidence that she receives in nearby cases, then the agent may be able to retain her knowledge, no matter how high the practical stakes are. In this respect, this view parts ways with standard accounts of pragmatic encroachment, which at least seemingly subscribe to the simple formula that *high practical stakes destroy knowledge.* On this view, high stakes can destroy knowledge only if they pose a risk of rational defeat. Second, this view remains silent about the connection between the practical interests and propositional justification: the practical interests of an agent can encroach on knowledge only by posing a risk of rational defeat, but not by affecting what counts as knowledge-level evidence in the relevant scenario. This account, therefore, is compatible with evidentialism; it does not force us to think that propositional justification is subject to any kind of pragmatic constraint.

3.7 Conclusion

In this chapter, I have explored a stability condition on knowledge: SAFETY FROM DEFEAT. This condition gives a unified explanation of a variety of epistemic phenomena: it tells us how knowledge explains rational perseverance, how we don't need certain 'internalist" conditions on knowledge, and how practical interests can encroach on knowledge. Thus, SAFETY FROM DEFEAT helps us make progress not only in *anti-luck* epistemology—the post-Gettier project of finding anti-luck conditions on knowledge—but also in the broader enterprise of discovering structural features of knowledge.

References

[1] Jonathan Adler. *Belief's Own Ethics.* Cambridge, MA: MIT Press, 2002.

[2] Jonathan E. Adler. Epistemics and the Total Evidence Requirement. *Philosophia,* 19(2-3):227–243, 1989.

[3] William P. Alston. An Internalist Externalism. *Synthese,* 74(3):265–283, 1988.

[4] D. M. Armstrong. *Belief, Truth and Knowledge,* volume 24. London: Cambridge University Press, 1973.

[5] D. M. Armstrong. Going Through the Open Door Again: Counterfactual Versus Singularist Theories of Causation. In Gerhard Preyer, editor, *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis,* pages 163–176. Rowman and Littlefield, 2001.

[6] Jerrold L. Aronson. On the Grammar of 'Cause'. *Synthese,* 22(3-4):414–430, 1971.

[7] Nomy Arpaly. *Unprincipled Virtue: An Inquiry Into Moral Agency.* Oxford: Oxford University Press, 2003.

[8] Max Baker-Hytch and Matthew A. Benton. Defeatism Defeated. *Philosophical Perspectives,* 29(1):40–66, 2015.

[9] Marcia Baron. *Kantian Ethics Almost Without Apology.* Ithaca: Cornell University Press, 1995.

[10] Helen Beebee. Causing and Nothingness. In L. A. Paul, E. J. Hall, and J. Collins, editors, *Causation and Counterfactuals,* pages 291–308. The MIT Press, 2004.

[11] Michael Bergmann. *Justification Without Awareness: A Defense of Epistemic Externalism*. Oxford: Oxford University Press, 2006.

[12] Selim Berker. The Rejection of Epistemic Consequentialism. *Philosophical Issues*, 23(1):363–387, 2013.

[13] Dylan Bianchi. Know-How and Information Access. ms. Unpublished Manuscript.

[14] Brand Blanshard. *Reason and Belief*. New Haven: Yale University Press, 1974.

[15] Laurence BonJour. *The Structure of Empirical Knowledge*. Cambridge: Cambridge University Press, 1985.

[16] Sylvain Bromberger. *On What We Know We Don't Know: Explanation, Theory, Linguistics, and How Questions Shape Them*. Chicago: University of Chicago Press, 1992.

[17] John Broome. Reasons. In R. Jay Wallace, editor, *Reason and Value: Themes From the Moral Philosophy of Joseph Raz*, pages 28–55. Oxford: Oxford University Press, 2004.

[18] John Broome. *Rationality Through Reasoning*. Wiley-Blackwell, 2013.

[19] Lara Buchak. Instrumental rationality, epistemic rationality, and evidence-gathering. *Philosophical Perspectives*, 24(1):85–120, 2010.

[20] Tyler Burge. Individualism and the Mental. *Midwest Studies in Philosophy*, 4(1):73–122, 1979.

[21] Tyler Burge. Individualism and Psychology. *Philosophical Review*, 95(1): 3–45, 1986.

[22] Fabrizio Cariani, Magdalena Kaufmann, and Stefan Kaufmann. Deliberative Modality Under Epistemic Uncertainty. *Linguistics and Philosophy*, 36 (3):225–259, 2013.

[23] Rudolf Carnap. *Logical Foundations of Probability*. Chicago: University of Chicago Press, 1962.

[24] E. F. Carritt. *Ethical and Political Thinking*, volume 26. Westport, Conn.,Greenwood Press, 1947.

[25] Roderick M Chisholm. *Theory of knowledge.* Englewood Cliffs, NJ: Prentice Hall, 1989. Third Edition.

[26] David Christensen. Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals. *Oxford Studies in Epistemology,* 2:3–31, 2007.

[27] David Christensen. Epistemic Self-Respect. *Proceedings of the Aristotelian Society,* 107(1pt3):319–337, 2007.

[28] David Christensen. Epistemology of Disagreement: The Good News. *Philosophical Review,* 116(2):187–217, 2007.

[29] David Christensen. Disagreement as Evidence: The Epistemology of Controversy. *Philosophy Compass,* 4(5):756–767, 2009.

[30] David Christensen. Higher-Order Evidence. *Philosophy and Phenomenological Research,* 81(1):185–215, 2010.

[31] David Christensen. Rational Reflection. *Philosophical Perspectives,* 24(1): 121–140, 2010.

[32] David Christensen. Higher-order evidence. *Philosophy and Phenomenological Research,* 81(1):185–215, 2010. ISSN 1933-1592. doi: 10.1111/j.1933-1592.2010.00366.x. URL `http://dx.doi.org/10.1111/j.1933-1592.2010.00366.x`.

[33] Ivano Ciardelli, Jeroen Groenendijk, and Floris Roelofsen. Attention! 'Might' in Inquisitive Semantics. In Satoshi Ito Ed Cormany and David Lutz, editors, *Proceedings of the 19th Semantics and Linguistic Theory Conference,* pages 91–108, 2009.

[34] E. Conee and R. Feldman. The Generality Problem for Reliabilism. *Philosophical Studies,* 89(1):1–29, 1998.

[35] Earl Conee. Heeding Misleading Evidence. *Philosophical Studies,* 103(2): 99–120, 2001.

[36] Earl Conee and Richard Feldman. *Evidentialism.* Oxford: Oxford University Press, 2004.

[37] Jonathan Dancy. *Practical Reality.* Oxford: Oxford University Press, 2002.

[38] Nilanjan Das and Bernhard Salow. Transparency and the KK Principle. *Noûs*, forthcoming.

[39] David Davidson. Paradoxes of Irrationality. In Richard Wollheim and James Hopkins, editors, *Philosophical Essays on Freud*, pages 289–305. Cambridge: Cambridge University Press, 1982.

[40] Donald Davidson. Actions, Reasons, and Causes. *Journal of Philosophy*, 60 (23):685–700, 1963.

[41] Donald Davidson. Deception and Division. In *Problems of Rationality*. Oxford: Oxford University Press, 2004.

[42] Eddie Dekel, Barton L Lipman, and Aldo Rustichini. Standard State-space Models Preclude Unawareness. *Econometrica*, pages 159–173, 1998.

[43] Keith DeRose. Contextualism and Knowledge Attributions. *Philosophy and Phenomenological Research*, 52(4):913–929, 1992.

[44] Keith DeRose. Solving the Skeptical Problem. *Philosophical Review*, 104 (1):1–52, 1995.

[45] Phil Dowe. A Counterfactual Theory of Prevention and 'Causation' by Omission. *Australasian Journal of Philosophy*, 79(2):216–226, 2001.

[46] Phil Dowe. Causes Are Physically Connected to Their Effects: Why Preventers and Omissions Are Not Causes. In Christopher Hitchcock, editor, *Contemporary Debates in Philosophy of Science*, pages 189–196. Oxford: Basil Blackwell, 2004.

[47] Fred Dretske. Epistemic Operators. *Journal of Philosophy*, 67(24):1007–1023, 1970.

[48] Fred Dretske. Conclusive reasons. *Australasian Journal of Philosophy*, 49 (1):1–22, 1971.

[49] John Earman. *Bayes or Bust?* Cambridge, MA: MIT Press, 1992.

[50] Andy Egan. Seeing and Believing: Perception, Belief Formation and the Divided Mind. *Philosophical Studies*, 140(1):47–63, 2008.

[51] Andy Egan. Comments on Gendler's, "The Epistemic Costs of Implicit Bias". *Philosophical Studies*, 156(1):65–79, 2011.

[52] Adam Elga. Reflection and Disagreement. *Noûs*, 41(3):478–502, 2007.

[53] Adam Elga and Agustin Rayo. Fragmentation and Information Access. ms. Unpublished Manuscript.

[54] Mylan Engel. Is Epistemic Luck Compatible with Knowledge? *Southern Journal of Philosophy*, 30(2):59??75, 1992.

[55] Jeremy Fantl and Matthew McGrath. Evidence, Pragmatics, and Justification. *Philosophical Review*, 111(1):67–94, 2002.

[56] Jeremy Fantl and Matthew Mcgrath. On Pragmatic Encroachment in Epistemology. *Philosophy and Phenomenological Research*, 75(3):558–589, 2007.

[57] Jeremy Fantl and Matthew McGrath. *Knowledge in an Uncertain World.* Oxford University Press, 2009.

[58] Richard Feldman. Respecting the Evidence. *Philosophical Perspectives*, 19 (1):95–119, 2005.

[59] Richard Feldman and Earl Conee. Evidentialism. *Philosophical Studies*, 48 (1):15–34, 1985.

[60] Hartry Field. Apriority as an Evaluative Notion. In Paul Boghossian and Christopher Peacocke, editors, *New Essays on the A Priori*, pages 117–149. Oxford: Clarendon Press, 2000.

[61] Roderick Firth. Epistemic Merit, Intrinsic and Instrumental. *Proceedings and Addresses of the American Philosophical Association*, 55(1):5–23, 1981.

[62] William Fish. *Perception, Hallucination, and Illusion.* Oxford University Press, 2009.

[63] Jerry A. Fodor. *The Modularity of Mind.* Cambridge, MA: MIT Press, 1983.

[64] Jerry A. Fodor. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, MA: MIT Press, 1987.

[65] Richard Foley. *When is True Belief Knowledge?* Princeton, NJ: Princeton University Press, 2012.

[66] J. Dmitri Gallow. How to Learn From Theory-Dependent Evidence; or Commutativity and Holism: A Solution for Conditionalizers. *British Journal for the Philosophy of Science*, 65(3):493–519, 2014.

[67] Ala Garfinkel. *Forms of Explanation*. Yale University Press New Haven, 1981.

[68] John Geanakoplos. Game Theory without Partitions, and Applications to Speculation and Consensus. Technical report, 1989. Cowles Foundation Discussion Paper No. 914, Yale University.

[69] Tamar Szabó Gendler. On the Epistemic Costs of Implicit Bias. *Philosophical Studies*, 156(1):33–63, 2011.

[70] Edmund Gettier. Is Justified True Belief Knowledge? *Analysis*, 23(6):121–123, 1963.

[71] Alan H Goldman. *Empirical knowledge*. Berkeley: University of California Press, 1988.

[72] Alvin Goldman. Discrimination and Perceptual Knowledge. *Journal of Philosophy*, 73(November):771–791, 1976.

[73] Alvin Goldman. Williamson on knowledge and evidence. In Patrick Greenough and Duncan Pritchard, editors, *Williamson on Knowledge*, pages 73–92. Oxford: Oxford University Press, 2009.

[74] Alvin Goldman. Reliabilism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2011 edition, 2011.

[75] Alvin I. Goldman. What is Justified Belief? In George Pappas, editor, *Justification and Knowledge*, pages 1–25. Boston: D. Reidel, 1979.

[76] Alvin I. Goldman. Review of *Philosophical Explanations* by Robert Nozick. *The Philosophical Review*, 92(1):81–88, 1983.

[77] Alvin I. Goldman. *Epistemology and Cognition*. Cambridge, MA: Harvard University Press, 1986.

[78] I. J. Good. On the Principle of Total Evidence. *British Journal for the Philosophy of Science*, 17(4):319–321, 1967.

[79] Melvyn Goodale and A. D. Milner. One Brain - Two Visual Systems. *Psychologist*, 19(11):660–663, 2006.

[80] Hilary Greaves. Epistemic Decision Theory. *Mind*, 122(488):915–952, 2013.

[81] Hilary Greaves and David Wallace. Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind*, 115(459): 607–632, 2006.

[82] Daniel Greco. A Puzzle About Epistemic Akrasia. *Philosophical Studies*, 167 (2):201–219, 2014.

[83] Daniel Greco. Could KK Be OK? *Journal of Philosophy*, 111(4):169–197, 2014.

[84] Daniel Greco. Iteration and Fragmentation. *Philosophy and Phenomenological Research*, 91(3):656–673, 2014.

[85] Daniel Greco. Cognitive Mobile Homes. *Mind*, forthcoming.

[86] Sven Ove Hansson. *A Textbook of Belief Dynamics: Theory Change and Database Updating*, volume 11. Dodrecht, Netherlands: Springer, 1999.

[87] Gilbert Harman. *Thought*. Princeton: Princeton University Press, 1973.

[88] John Hawthorne. *Knowledge and Lotteries*. Oxford: Oxford University Press, 2004.

[89] John Hawthorne. A Priority and Externalism. In Sanford Goldberg, editor, *Internalism and Externalism in Semantics and Epistemology*, pages 201–218. Oxford: Oxford University Press, 2007.

[90] John Hawthorne and Amia Srinivasan. Disagreement Without Transparency: Some Bleak Thoughts. In David Christensen and Jennifer Lackey, editors, *The Epistemology of Disagreement: New Essays*, pages 9–30. Oxford University Press, 2013.

[91] John Hawthorne and Jason Stanley. Knowledge and Action. *Journal of Philosophy*, 105(10):571–590, 2008.

[92] Carl Hempel. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, 1965.

[93] Barbara Herman. On the Value of Acting From the Motive of Duty. *Philosophical Review*, 90(3):359–382, 1981.

[94] Stephen Hetherington. Actually Knowing. *Philosophical Quarterly*, 48 (193):453–469, 1998.

[95] C. Hitchcock. Contrastive Explanation and the Demons of Determinism. *British Journal for the Philosophy of Science*, 50(4):585–612, 1999.

[96] Sophie Horowitz. Epistemic Akrasia. *Noûs*, 48(4):718–744, 2014.

[97] David Hume. *An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press, 2000/1748.

[98] Rosalind Hursthouse. *On Virtue Ethics*. Oxford: Oxford University Press, 1999.

[99] C. S. Jenkins. Knowledge and Explanation. *Canadian Journal of Philosophy*, 36(2):137–163, 2006.

[100] C. S. Jenkins. Entitlement and Rationality. *Synthese*, 157(1):25–45, 2007.

[101] Mark Johnston. Better Than Mere Knowledge? The Function of Sensory Awareness. In *Perceptual Experience*, pages 260–290. Oxford: Oxford University Press, 2006.

[102] James Joyce. Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In Franz Huber and Christoph Schmidt-Petri, editors, *Degrees of Belief*, pages 263–297. Synthese, 2009.

[103] James M Joyce. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, pages 575–603, 1998.

[104] Joseph B. Kadane, Mark Schervish, and Teddy Seidenfield. Is Ignorance Bliss? *Journal of Philosophy*, 105(1):5–36, 2008.

[105] Immanuel Kant. *Groundwork for the Metaphysics of Morals*. Cambridge: Cambridge University Press, 2012/1785. Translated and edited by Mary Gregor and Jens Timmerman.

[106] Thomas Kelly. The Epistemic Significance of Disagreement. In John Hawthorne and Tamar Gendler, editors, *Oxford Studies in Epistemology, Volume 1*, volume 1, pages 167–196. Oxford: Oxford University Press, 2005.

[107] Thomas Kelly. Disagreement, Dogmatism, and Belief Polarization. *Journal of Philosophy*, 105(10):611–633, 2008.

[108] Thomas Kelly. Evidence. In *Stanford Encyclopedia of Philosophy*, volume 3, pages 933–955. 2008.

[109] Thomas Kelly. Evidence: Fundamental Concepts and the Phenomenal Conception. *Philosophy Compass*, 3(5):933–955, 2008.

[110] Thomas Kelly. Peer Disagreement and Higher Order Evidence. In Alvin I. Goldman and Dennis Whitcomb, editors, *Social Epistemology: Essential Readings*, pages 183–217. Oxford University Press, 2010.

[111] Peter D. Klein. A Proposed Definition of Propositional Knowledge. *Journal of Philosophy*, 68(16):471–482, 1971.

[112] Peter D. Klein. *Certainty, a Refutation of Scepticism*. Minneapolis: University of Minnesota Press, 1981.

[113] Niko Kolodny and John MacFarlane. Ifs and Oughts. *Journal of Philosophy*, 107(3):115–143, 2010.

[114] Matthew Kotzen. Selection Biases in Likelihood Arguments. *British Journal for the Philosophy of Science*, 63(4):825–839, 2012.

[115] Saul A. Kripke. Nozick on knowledge. In *Philosophical Troubles. Collected Papers Vol I.* Oxford: Oxford University Press, 2011.

[116] Saul A. Kripke. Two paradoxes of knowledge. In *Philosophical Troubles. Collected Papers Vol I.* Oxford: Oxford University Press, 2011.

[117] Maria Lasonen-Aarnio. Unreasonable Knowledge. *Philosophical Perspectives*, 24(1):1–21, 2010.

[118] Maria Lasonen-Aarnio. The Dogmatism Puzzle. *Australasian Journal of Philosophy*, (3):1–16, 2013.

[119] Maria Lasonen-Aarnio. Higher-Order Evidence and the Limits of Defeat. *Philosophy and Phenomenological Research*, 88(2):314–345, 2014.

[120] Maria Lasonen-Aarnio. New Rational Reflection and Internalism about Rationality. *Oxford Studies in Epistemology*, 5:145, 2015.

[121] Maria Lasonen-Aarnio. Enkrasia or Evidentialism? Learning to Love Mismatch. ms. Unpublished manuscript.

[122] Keith Lehrer. Knowledge, Truth and Evidence. *Analysis*, 25(5):168–175, 1965.

[123] Keith Lehrer. *Knowledge*. Oxford: Clarendon Press, 1974.

[124] Keith Lehrer. *Theory of Knowledge*. San Francisco and Boulder: Westview Press, 1990.

[125] Keith Lehrer and Thomas Paxson Jr. Knowledge: Undefeated justified true belief. *Journal of Philosophy*, 66(8):225–237, 1969.

[126] Hannes Leitgeb and Richard Pettigrew. An Objective Justification of Bayesianism I: Measuring Inaccuracy. *Philosophy of Science*, 77(2):201–235, 2010.

[127] Hannes Leitgeb and Richard Pettigrew. An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science*, 77(2):236–272, 2010.

[128] David Lewis. Counterfactual Dependence and Time's Arrow. *Noûs*, 13(4): 455–476, 1979.

[129] David Lewis. Logic for Equivocators. *Noûs*, 16(3):431–441, 1982.

[130] David Lewis. Elusive Knowledge. *Australasian Journal of Philosophy*, 74(4): 549–567, 1996.

[131] David Lewis. Void and Object. In John Collins, Ned Hall, and L. A. Paul, editors, *Causation and Counterfactuals*, pages 277–290. MIT Press, 2004.

[132] Clayton Littlejohn. *Justification and the Truth-Connection*. Cambridge: Cambridge University Press, 2012.

[133] Clayton Littlejohn. Stop Making Sense? A Puzzle About Epistemic Rationality. *Philosophy and Phenomenological Research*, forthcoming.

[134] Steven Luper. False Negatives. In Kelly Becker and Tim Black, editors, *The sensitivity principle in epistemology*. Cambridge: Cambridge University Press, 2012.

[135] William G. Lycan. Evidence One Does Not Possess. *Australasian Journal of Philosophy*, 55(2):114–126, 1977.

[136] Alex Madva. Virtue, Social Knowledge, and Implicit Bias. In Jennifer Saul and Michael Brownstein, editors, *Implicit Bias and Philosophy*. Oxford: Oxford University Press, forthcoming.

[137] P. D. Magnus and Jonathan Cohen. Williamson on Knowledge and Psychological Explanation. *Philosophical Studies*, 116(1):37–52, 2003.

[138] Patrick Maher. Subjective and Objective Confirmation. *Philosophy of Science*, 63(2):149–174, 1996.

[139] Julia Markovits. Acting for the Right Reasons. *Philosophical Review*, 119 (2):201–242, 2010.

[140] Julia Markovits. *Moral Reason*. Oxford: Oxford University Press, 2014.

[141] Jack Marley-Payne. Task-Indexed Belief. ms. Unpublished Manuscript.

[142] Lisa Miracchi. Competence to Know. *Philosophical Studies*, 172(1):29–56, 2015.

[143] Salvatore Modica and Aldo Rustichini. Awareness and Partitional Information Structures. *Theory and Decision*, 37(1):107–124, 1994.

[144] Bernard Molyneux. Primeness, Internalism and Explanatory Generality. *Philosophical Studies*, 135(2):255–277, 2007.

[145] Jennifer Nagel. The Psychological Basis of the Harman-Vogel Paradox. *Philosophers' Imprint*, 11(5):1–28, 2011.

[146] Jennifer Nagel. Knowledge as a Mental State. *Oxford Studies in Epistemology*, 4:275–310, 2013.

[147] Ram Neta and Guy Rohrbaugh. Luminosity and the safety of knowledge. *Pacific Philosophical Quarterly*, 85(4):396–406, 2004.

[148] Robert Nozick. *Philosophical Explanations*. Cambridge, MA: Harvard University Press, 1981.

[149] Robert Nozick. *Socratic Puzzles*. Cambridge, MA: Harvard University Press, 1997.

[150] David Owens. *Causes and Coincidences.* Cambridge: Cambridge University Press, 1992.

[151] Derek Parfit. What We Together Do. ms.

[152] J.L. Pollock and J. Cruz. *Contemporary Theories of Knowledge.* Lanham and Oxford: Rowman & Littlefield, 1999. Second Edition.

[153] Harold Arthur Prichard. *Duty and Ignorance of Fact: From the Proceedings of the British Academy.* London: Humphrey Milford, 1932.

[154] Duncan Pritchard. *Epistemic Luck.* Oxford: Clarendon Press, 2005.

[155] James Pryor. Highlights of Recent Epistemology. *British Journal for the Philosophy of Science,* 52(1):95–124, 2001.

[156] James Pryor. Problems for Credulism. In Chris Tucker, editor, *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism.* Oxford: Oxford University Press, 2013.

[157] Hilary Putnam. The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science,* 7:131–193, 1975.

[158] Colin Radford. Knowledge—By Examples. *Analysis,* 27(1):1–11, 1966.

[159] Peter Railton. Probability, Explanation, and Information. *Synthese,* 48(2): 233–256, 1981.

[160] Agustin Rayo. *The Construction of Logical Space.* Oxford: Oxford University Press, 2013.

[161] Steven Rieber. Skepticism and Contrastive Explanation. *Noûs,* 32(2):189–204, 1998.

[162] Susanna Rinard. No Exception for Belief. *Philosophy and Phenomenological Research,* 91(2), 2015.

[163] David Ross. *Foundations of Ethics.* Oxford University Press, 1939.

[164] Bertrand Russell. The Elements of Ethics. In *Philosophical Essays.* London: Longmans, Green and Co., 1910.

[165] Bertrand Russell. *Human Knowledge: Its Scope and Limits.* London: Routledge, 2009/1948.

[166] Wesley C. Salmon. *The Foundations of Scientific Inference.* Pittsburgh: University of Pittsburgh Press, 1967.

[167] Jonathan Schaffer. Causation by Disconnection. *Philosophy of Science,* 67 (2):285–300, 2000.

[168] Jonathan Schaffer. Causes Need Not Be Physically Connected to Their Effects: The Case for Negative Causation. In Christopher Read Hitchcock, editor, *Contemporary Debates in Philosophy of Science,* pages 197–216. Oxford: Basil Blackwell, 2004.

[169] Joshua Schechter. Rational Self-Doubt and the Failure of Closure. *Philosophical Studies,* 163(2):428–452, 2013.

[170] Miriam Schoenfield. Permission to Believe: Why Permissivism Is True and What It Tells Us About Irrelevant Influences on Belief. *Noûs,* 47(1):193–218, 2013.

[171] Miriam Schoenfield. A Dilemma for Calibrationism. *Philosophy and Phenomenological Research,* 91(2):425–455, 2015.

[172] Eric Schwitzgebel. Acting Contrary to Our Professed Beliefs or the Gulf Between Occurrent Judgment and Dispositional Belief. *Pacific Philosophical Quarterly,* 91(4):531–553, 2010.

[173] Kieran Setiya. *Knowing Right From Wrong.* Oxford: Oxford University Press, 2012.

[174] Reza Shadmehr and Ferdinando A Mussa-Ivaldi. Adaptive Representation of Dynamics during Learning of a Motor Task. *The Journal of Neuroscience,* 14(5):3208–3224, 1994.

[175] Brian Skyrms. The Value of Knowledge. *Minnesota Studies in the Philosophy of Science,* 14:245–266, 1990.

[176] Paulina Sliwa and Sophie Horowitz. Respecting All the Evidence. *Philosophical Studies,* 172(11):2835–2858, 2015.

[177] Adam Smith and Knud Haakonssen. *Adam Smith: The Theory of Moral Sentiments.* Cambridge: Cambridge University Press, 2002/1759.

[178] Holly M. Smith. The Moral Clout of Reasonable Beliefs. In Mark Timmons, editor, *Oxford Studies in Normative Ethics*, volume 1. Oxford University Press, 2010.

[179] Michael Smith. The Humean Theory of Motivation. *Mind*, 96(381):36–61, 1987.

[180] Declan Smithies. Moore's Paradox and the Accessibility of Justification. *Philosophy and Phenomenological Research*, 85(2):273–300, 2012.

[181] Declan Smithies. Ideal Rationality and Logical Omniscience. *Synthese*, 192 (9):2769–2793, 2015.

[182] Declan Smithies. Reflection On: On Reflection. *Analysis*, forthcoming.

[183] Declan Smithies. Why Justification Matters. In David Henderson and John Greco, editors, *Epistemic Evaluation: Point and Purpose in Epistemology*. Oxford: Oxford University Press, ms.

[184] Elliott Sober. *Simplicity*. Clarendon Press, 1975.

[185] Elliott Sober. Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads. *Philosophical Studies*, 143(1):63–90, 2009.

[186] Richard Sorabji. *Necessity, Cause, and Blame: Perspectives on Aristotle's Theory*. University of Chicago Press, 1980.

[187] Ernest Sosa. How Must Knowledge Be Modally Related to What Is Known? *Philosophical Topics*, 26(1/2):373–384, 1999.

[188] Ernest Sosa. Tracking, Competence, and Knowledge. In Paul K. Moser, editor, *The Oxford Handbook of Epistemology*, pages 264–287. Oxford: Oxford University Press, 2002.

[189] Amia Srinivasan. Normativity Without Cartesian Privilege. *Philosophical Issues*, 25(1):273–299, 2015.

[190] Robert Stalnaker. *Inquiry*. Cambridge, MA: MIT Press, 1984.

[191] Robert Stalnaker. The Problem of Logical Omniscience, I. *Synthese*, 89(3): 425–440, 1991.

[192] Robert Stalnaker. Luminosity and the KK Thesis. In Sanford C. Goldberg, editor, *Externalism, Self-Knowledge, and Skepticism*, volume 1, pages 167–196. Cambridge: Cambridge University Press, 2015.

[193] Jason Stanley. *Knowledge and Practical Interests.* Oxford: Oxford University Press, 2005.

[194] Gail Stine. Skepticism, relevant alternatives, and deductive closure. *Philosophical Studies*, 29(4):249–261, 1976.

[195] Philip Stratton-Lake. *Kant, Duty, and Moral Worth.* London: Routledge, 2000.

[196] Jonathan Sutton. *Without Justification.* Cambridge, MA: MIT Press, 2007.

[197] Marshall Swain. Epistemic Defeasibility. *American Philosophical Quarterly*, 11(1):15–25, 1974.

[198] Judith Jarvis Thomson. *Rights, Restitution, and Risk: Essays, in Moral Theory.* Cambridge, MA: Harvard University Press, 1986.

[199] Michael G Titelbaum. Rationality's Fixed Point (or: In Defense of Right Reason). *Oxford Studies in Epistemology*, 5:253, 2015.

[200] Endel Tulving and Donald M Thomson. Encoding Specificity and Retrieval Processes in Episodic Memory. *Psychological review*, 80(5):352, 1973.

[201] John Turri. In Gettier's wake. In Stephen Hetherington, editor, *Epistemology: The Key Thinkers*, pages 214–229. New York: Continuum, 2012.

[202] Peter Unger. An Analysis of Factual Knowledge. *Journal of Philosophy*, 65 (6):157–170, 1968.

[203] B. C. Van Fraassen. *The Scientific Image.* Oxford: Clarendon Press, 1980.

[204] Han van Wietmarschen. Peer Disagreement, Evidence, and Well-Groundedness. *Philosophical Review*, 122(3):395–425, 2013.

[205] Jonathan Vogel. Tracking, Closure, and Inductive Knowledge. In Luper-Foy Steven, editor, *The Possibility of Knowledge: Nozick and His Critics*, pages 197–215. Rowman & Littlefield, 1987.

[206] Brian Weatherson. Can We Do Without Pragmatic Encroachment? *Philosophical Perspectives*, 19(1):417–443, 2005.

[207] Brian Weatherson. Knowledge, Bets, and Interests. In Jessica Brown and Mikkel Gerken, editors, *Knowledge Ascriptions*, pages 75–103. Oxford: Oxford University Press, 2012.

[208] Ralph Wedgwood. Justified Inference. *Synthese*, 189(2):1–23, 2012.

[209] Jonathan Weisberg. Commutativity or Holism? A Dilemma for Conditionalizers. *British Journal for the Philosophy of Science*, 60(4):793–812, 2009.

[210] Jonathan Weisberg. Updating, Undermining, and Independence. *British Journal for the Philosophy of Science*, 66(1):121–159, 2015.

[211] Roger White. Problems for Dogmatism. *Philosophical Studies*, 131(3):525–57, 2006.

[212] John N. Williams. Not Knowing You Know: A New Objection to the Defeasibility Theory of Knowledge. *Analysis*, 75(2):213–217, 2015.

[213] Timothy Williamson. *Knowledge and its Limits*. Oxford: Oxford University Press, 2000.

[214] Timothy Williamson. Reply to Goldman. In Patrick Greenough and Duncan Pritchard, editors, *Williamson on Knowledge*. Oxford: Oxford University Press, 2009.

[215] Timothy Williamson. Reply to John Hawthorne and Maria Lasonen-Aarnio. In Patrick Greenough and Duncan Pritchard, editors, *Williamson on Knowledge*. Oxford: Oxford University Press, 2009.

[216] Timothy Williamson. Improbable Knowing. In Trent Dougherty, editor, *Evidentialism and its Discontents*. Oxford University Press, 2011.

[217] Alex Worsnip. The Conflict of Evidence and Coherence. *Philosophy and Phenomenological Research*, forthcoming.

[218] Crispin Wright. Warrant for Nothing (and Foundations for Free)? *Aristotelian Society Supplementary Volume*, 78(1):167–212, 2004.

[219] Seth Yalcin. Nonfactualism About Epistemic Modality. In *Epistemic Modality*. Oxford: Oxford University Press, 2011.