Forthcoming in *The Journal of Philosophy*. (A few minor revisions still to come.)

Disagreement, Dogmatism, and Belief Polarization

Thomas Kelly
Princeton University

The human mind gets creased into a way of seeing things.

--Antoine Lavoisier, *Reflections on Phlogiston*

1. Introduction

Consider the phenomenon of **belief polarization**. Suppose that two individuals—let's call them 'You' and 'I'--disagree about some non-straightforward matter of fact: say, about whether capital punishment tends to have a deterrent effect on the commission of murder. Although neither of us is certain of his or her view, I believe that capital punishment is a deterrent while You believe that it is not. Perhaps one or both of us has evidence for his or her view. Or perhaps we hold our views on the basis of ideological dogma, or on the basis of some admixture of dogma and evidence. In any case, regardless of why we believe as we do, You and I disagree, in a perfectly familiar way.[1]

Suppose next that the two of us are subsequently exposed to a relatively substantial body of evidence that bears on the disputed question: for example, statistical studies comparing the murder rates for adjacent states with and without the death penalty. The evidence is of a mixed character: some studies seem to suggest that capital punishment is a deterrent while other studies seem to suggest that it is not. Regardless, the entire body

[1] Here and throughout, I use 'disagree' in a weak sense, according to which you and I disagree about some issue just in case we hold opposed views about that issue. In particular, as I will use the term, it does not follow from the fact that you and I disagree that we are *aware* that we hold opposed views (or indeed, even that we are aware that the other exists at all). Questions about how we should respond to an awareness of disagreement are ones that I have pursued at some length elsewhere; see 'The Epistemic Significance of Disagreement' in Tamar Szabo Gendler and John Hawthorne (eds.) *Oxford Studies in Epistemology*, vol.1 (Oxford: Oxford University Press, 2005): 167-196, and 'Peer Disagreement and the "Common Consent" Argument for the Existence of God: the Views of Similarly Situated Others as Evidence' in Richard Feldman and Ted Warfield (eds.) *Disagreement* (forthcoming from Oxford University Press). But they will not be on the agenda here.

of evidence is presented to each of us: there is no piece of evidence that is available to you but not to me, or vice versa.

What becomes of our initial disagreement once we are exposed to such evidence? It is natural to expect—and perhaps, also natural to hope—that mutual exposure to common evidence will tend to lessen or mitigate our disagreement. Perhaps it would be unrealistic to expect a perfect convergence of opinion: after all, we begin with diametrically opposed views, and one might expect this fact to find reflection in our later opinions. Still, it's natural to expect that our exposure to common evidence will tend to narrow the gap between us and that, indeed, as the total evidence which is available to each of us increasingly comes to consist of common items, our views will undergo a corresponding convergence. (A Bayesian might speak here of the 'swamping' or 'washing out' of our respective prior probabilities.) At the very least, one would expect that exposure to common evidence would not *increase* the extent of our disagreement.

In fact, however, if You and I are typical of subjects who have participated in actual experiments of exactly this sort, such natural expectations will be disappointed.[2] Exposure to evidence of a mixed character does *not* typically narrow the gap between those who hold opposed views at the outset. Indeed, worse still: not only is convergence typically not forthcoming, but in fact, exposure to such evidence tends to make initial disagreements even more pronounced. The more I am exposed to evidence of a mixed character, the more confident I tend to become of my view that capital punishment is a deterrent. On the other hand, the more You are exposed to the same evidence, the more confident You tend to become of your initial view that capital punishment is *not* a deterrent. As our shared evidence increases, each of us tends to harden in his or her opinion, and the gulf between us widens. Our attitudes become increasingly polarized.[3]

---

[2] The classic study in this area is Charles Lord, Lee Ross, and Mark Lepper, 'Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence', *Journal of Personality and Social Psychology*, 37, 11 (1979): 2098-2109, from which the example of the death penalty is taken. A useful overview of relevant literature, including follow-up studies, is Thomas Gilovich *How We Know What Isn't So* (New York: The Free Press, 1991); see especially Chapter 3, 'Seeing what we expect to see'.

[3] The results in question are thus perhaps especially disappointing from the perspective of a certain kind of (old-fashioned?) Enlightenment line of thought, a line of thought which runs as follows. The fact that individuals disagree so strongly about various controversial normative issues (e.g., whether the death penalty ought to be abolished, whether institutions should have strong affirmative action policies) is often

The empirically well-confirmed phenomenon of belief polarization is, I think, an interesting and potentially important one, and one that it is worth attempting to understand better. What, exactly, are You and I doing? (Or not doing?) There are two sets of questions here. The first set consists of purely descriptive, psychological questions about how exactly You and I are responding to our evidence so as to generate the relevant phenomenon. The second set consists of normative questions. Given that You and I are responding to our evidence in such-and-such a way, is there any chance that our doing so is anything other than blatantly unreasonable? What is the epistemic status of the views at which we ultimately arrive by responding to our evidence in this way? How (if at all) should we attempt to counteract or correct for the relevant psychological tendency? As we will see, these normative questions are less straightforward than one might expect; pursuing them raises a number of rather subtle and delicate issues about what is to be objective or open-minded (on the one hand) as opposed to dogmatic or biased (on the other), as well as questions about the role that one's background beliefs should and should not play in the assessment of new evidence.

Although my ultimate concern is with the normative questions, I begin by attending to the psychological ones.

2. Kripkean Dogmatism

How then are You and I responding to the evidence with which we are presented? One possibility is the following: You and I are dogmatists, in something like the sense of Saul Kripke's 'Dogmatism Paradox'.[4] At the outset, I believe

---

underwritten by a kind of purely factual, non-normative ignorance on the part of one or both parties to the dispute (e.g., as to whether the death penalty functions as an effective deterrent, or what the true effects of affirmative action policies are). Thus, the key to eliminating (or at least, damping down) such normative disagreement is to do more and better social science, and then to make sure that the results of the relevant inquiries are adequately disseminated within society. The phenomenon of polarization should, I think, give at least some pause to one attracted to this general picture.

[4] Kripke, 'On Two Paradoxes of Knowledge', unpublished lecture delivered to the Cambridge Moral Sciences Club. The first published discussion of the paradox is Gilbert Harman, *Thought* (Princeton: Princeton University Press, 1973): 148-149. What follows is in fact a variant on Kripke's original puzzle, inasmuch as it employs the concept of justified belief rather than the concept of knowledge.

DETERRENCE: Capital punishment is a deterrent.

It follows immediately from DETERRENCE that

> Any evidence that suggests that capital punishment is not a deterrent is misleading evidence.

But of course, if I am concerned to believe the truth about a given question, then a policy of ignoring misleading evidence that bears on that question would seem to be a sensible policy to follow. Realizing this, I take my original belief that DETERRENCE is true as a license to dismiss evidence that suggests otherwise on the grounds that such evidence must be misleading. On the other hand, I see no such reason to dismiss subsequently encountered evidence that seems to support DETERRENCE. (Indeed, my belief that DETERRENCE is true might very well dispose me to expect that non-misleading evidence in its favor is likely to be forthcoming.) Thus, when I am exposed to a mixed body of evidence, I dismiss that portion which conflicts with my original belief while giving weight to that portion which supports it. As a result, I become increasingly confident that DETERRENCE is true. On the other hand, You reason in a parallel but opposite way, and thus become ever more confident that DETERRENCE is false. The net result is that You and I become increasingly polarized, as each of us treats his own belief as a license to discount exactly that portion of our shared evidence which, if taken at face value, would seem to support the other's point of view.[5]

If this is in fact an accurate description of how You and I are reasoning about our shared evidence, then You and I are unreasonable. That is: it is uncontroversial *that*

---

[5] Compare the anecdote related by C.S. Peirce in his classic essay 'The Fixation of Belief':

> …I remember once being entreated not to read a certain newspaper lest it might change my opinion upon free-trade. 'Lest I might be entrapped by its fallacies and misstatements,' was the form of expression. 'You are not', my friend said, 'a special student of political economy. You might, therefore, easily be deceived by fallacious arguments upon the subject. You might, then, if you read this paper, be led to believe in protection. But you admit that free-trade is the true doctrine; and you do not wish to believe what is not true'.

Peirce remarks that 'I have often known this system to be deliberately adopted'. (As reprinted in Justus Buchler (ed.) *Philosophical Writings of Peirce* (New York: Dover, 1955), p.11.)

Kripkean dogmatism is unreasonable. It is not immediately obvious *why* such reasoning is illegitimate, particularly if we consider cases in which one's belief is initially based on evidence sufficient to justify it (i.e., cases in which one possesses evidence sufficient to justify one's belief prior to being presented with the relevant statistical information). After all, if my original belief is justified at the outset, then, given a very plausible closure principle about justification[6], I am also justified in believing that any apparent counterevidence that I might encounter will be misleading. But if I am *justified* in believing that any counterevidence will be misleading, why am I not justified in ignoring such evidence when I actually encounter it?

This is a genuine philosophical puzzle. Fortunately, the solution to this particular puzzle is relatively well-understood, due to the work of philosophers such as Gilbert Harman and Roy Sorensen.[7] Because I have little to add to what these thinkers have said about the Dogmatism paradox, I will not linger over it here. I mention the possibility that You and I are Kripkean dogmatists in order to contrast it with an alternative descriptive model of how You and I are responding to evidence that seems to tell against our beliefs, a model that I take up in the next section. Before leaving the Dogmatism paradox however, I want to make use of it in order to introduce an idea that will be of some importance in what follows.

---

[6] The principle in question is that if S is justified in believing p, and S recognizes that p entails q, then S is justified in believing q.

[7] I take the essentials of the correct solution to have been provided by Harman *op. cit.*, p.149, in his original presentation of the puzzle, with much useful elaboration and further development provided by Sorensen, 'Dogmatism, Junk Knowledge, and Conditionals', *The Philosophical Quarterly* 38, 153 (1988):433-454.

Roughly, Harman's solution runs as follows. Suppose that at time t0 my total evidence is such as to justify my belief that p is true. Given a plausible closure principle about justification, I am thus also justified in believing, at time t0, that any subsequently encountered evidence against p will be misleading evidence. Why then, when I subsequently encounter evidence against p at time t1, am I not justified in concluding that it is misleading? Answer: Because once I encounter evidence against p at time t1, I may no longer be justified in believing that p is true, and (hence) no longer justified in believing that any evidence against p is misleading. There is thus no single time at which I both possess the evidence against p and am justified in concluding that that evidence is misleading, although any time at which I am justified in believing that p is true is also a time at which I am justified in believing that any evidence against p is misleading.

On the dogmatism paradox, see also Carl Ginet 'Knowing Less by Knowing More', *Midwest Studies in Philosophy* 5 (1980): 151-161; Tom Sorell, 'Harman's Paradox', *Mind*, New Series, 90, 360 (1981): 557-575; James Cargile, 'Justification and Misleading Defeaters', *Analysis* 55/3 (1995): 216-220; and Earl Conee, 'Heeding Misleading Evidence', *Philosophical Studies* 103 (2001): 99-120.

One route to appreciating the bankruptcy of Kripkean dogmatism is the following. Kripkean dogmatism would seem to allow facts about what one is justified in believing to depend in an implausible way on historical facts about the temporal order in which particular pieces of evidence are acquired. Suppose that at time t0 I have no opinion at all about whether some hypothesis H is true. (Perhaps I've simply never considered the matter before.) Although I have no opinion, I'm disposed to reason in the manner of a Kripkean dogmatist: as soon as I do form an opinion one way or the other, I will treat that opinion as a reason to dismiss any subsequently encountered evidence which seems to count against it. (Thus, I'm something of an *open-minded* Kripkean dogmatist: at the outset, I'm not wedded, or even disposed, to either believing or disbelieving the hypothesis in question.) Let **E1** be a piece of evidence that strongly confirms hypothesis H. Indeed, suppose that if E1 exhausted my total evidence with respect to H, then I would be justified in believing H on its basis. (Perhaps E1 is the testimony of a highly-reliable, though not infallible, authority that H is true.) Let **E2** be a piece of evidence that strongly disconfirms H; indeed, suppose that if E2 exhausted my total evidence with respect to H, then I would be justified in believing that *H is false* on its basis. (Perhaps E2 is the testimony of another, equally-reliable authority that H is false.)

Suppose that I am subsequently exposed to both E1 and E2 but to no other evidence that bears on H. Inasmuch as I am a Kripkean dogmatist, whether I end up believing that H is true or end up believing that H is false will depend crucially on the temporal order in which I encounter the two pieces of evidence. If I first encounter E1, I will acquire the justified belief that H is true and the justified belief that any evidence against H is misleading evidence; when I subsequently encounter E2, I will accordingly dismiss it as misleading and end up believing H. If, on the other hand, I encounter E2 first, then I will acquire the justified belief that H is false and the justified belief that any evidence which supports H is misleading; accordingly, when I subsequently encounter E1, I will dismiss *it* as misleading and end up believing that H is false. I thus end up with diametrically opposed views in the two cases, despite the fact that I have been exposed to exactly the same evidence in each. If we suppose that E1 and E2 consist of the conflicting testimony of two equally reliable authorities, what I end up believing will depend upon which of the two authorities I consulted first and which second. (Even if my decision of whom to

consult first was based on whose office happened to be closer to my home, or on the flip of a coin). Moreover, if I'm self-aware of my own practice, I would have knowledge of the following form: 'Because I came across evidence E1 before I came across evidence E2, I now believe that the hypothesis H is true. But if I had come across evidence E2 before evidence E1, I would now believe that H is false.'

It seems implausible (to say the least) that historical facts about the order in which evidence is acquired might make such a dramatic difference to what one is justified in believing. Indeed, many take it be a criterion of adequacy on any account of rational or justified belief that the order in which pieces of evidence are acquired makes *no difference at all* to what is reasonable for one to believe. This is the frequently endorsed requirement that evidence be commutative:

> **The Commutativity of Evidence Principle:** to the extent that what it is reasonable for one to believe depends on one's total evidence, historical facts about the order in which that evidence is acquired make no difference to what it is reasonable for one to believe.[8]

In what follows, I will assume that the Commutativity of Evidence Principle is true. As we will see, subtle epistemological issues can arise about how this Principle should be interpreted and applied in particular cases; ultimately, considerable refinement will be needed. For now, however, I want to turn to an alternative descriptive model of how individuals respond to evidence that seems to tell against their beliefs, a model which serves as a rival to Kripkean dogmatism as an account of the reasoning which underwrites the polarization phenomenon.

---

[8] Commitment to the principle is exhibited, for example, in the frequently made charge that Jeffrey conditionalization (as elaborated in Richard Jeffrey, *The Logic of Decision* (McGraw-Hill: New York, 1965)) fails to respect it and is for that reason inadequate. For this objection, see, among others, Frank Doring, 'Why Bayesian Psychology is Incomplete', *Philosophy of Science* 66 (Proceedings), S379-389; Brian Skyrms, *Choice and Chance*, 3rd ed. (Wadsworth: Belmont, CA, 1986), and Bas van Fraassen *Laws and Symmetry* (Clarendon Press: Oxford, 1989). Lange ('Is Jeffrey Conditionalization Defective By Virtue of Being Non-Commutative? Remarks on the Sameness of Sensory Experience' *Synthese* 123(2000):393-403) also accepts the principle but denies Jeffrey conditionalization runs afoul of it.

The principle is also sometimes endorsed by psychologists; see, for example Jonathan Baron, *Thinking and Deciding* (Cambridge: Cambridge University Press, 2000) p. 197.

3. An Alternative Model

It is characteristic of the Kripkean dogmatist to treat apparent counterevidence in a dismissive manner. Indeed, a Kripkean dogmatist need not even attend to the specific content of such evidence: as soon as he knows that a given piece of evidence tells against one of his beliefs, he knows all that he needs to know in order to employ his general policy; he thus pays it no further heed. The suggestion that You and I are Kripkean dogmatists is, no doubt, an unflattering one. I am thus happy to report that You and I do not seem to be dogmatists in this sense. That is, individuals who have participated in the relevant experiments do *not* typically pay less attention to counterevidence than to supporting evidence. Indeed, the opposite seems to be true: far from paying less attention to counterevidence, it seems that we pay *more* attention to it.[9]

Why would paying more attention to apparent counterevidence give rise to the polarization phenomenon? As a point of comparison, consider the way in which one's disbelieving the conclusion of an argument might play a role in one's uncovering a flaw in that argument--say, a subtle equivocation between the argument's premises and its conclusion. Typically, if one believes that p, then one also believes (or at least, is disposed to believe) that *there are no sound arguments for not-p*. When one is subsequently presented with what purports to be a sound argument for not-p, one is thus disposed to view that argument with a greater measure of suspicion and to subject it to closer scrutiny. And the more one subjects the argument to close scrutiny (roughly: the more cognitive resources one devotes to the task of finding some flaw in the argument), the more likely one is to find a flaw in that argument if in fact there is some flaw to be found. Of course, individuals can, and not infrequently do, recognize that particular arguments are flawed even when they agree with the conclusions of those arguments. But in general, there is evidence which suggests that our sensitivity to even formal fallacies is not invariant with respect to our prior attitude towards the conclusions of the arguments in which those fallacies are embedded. All else being equal, individuals tend to be significantly better at detecting fallacies when the fallacy occurs in an argument for

[9] See, for example, the discussion in Gilovich, *op. cit.*, chapter 3, especially pages 54-56.

a conclusion which they disbelieve, than when the same fallacy occurs in an argument for a conclusion which they believe.[10]

Suppose that one is presented with an argument for a conclusion which contradicts something that one believes. One examines the argument, judges that it is not a good one, and so retains one's original view. Suppose, moreover, that the fact that one judges that the argument is not a good one is contingent on the fact that one already disbelieved its conclusion prior to having been presented with the argument: if one had not already disbelieved the conclusion—if, say, one had been an agnostic or had not yet formed an opinion about the relevant issue—then one *would* have been persuaded by the argument. This might look extremely suspicious. After all, if an argument is sufficiently attractive that it would have convinced one if one had initially examined it from a standpoint of neutrality, how can it be legitimate for the crucial difference to be made by the fact that one *already* had an opinion about the issue in question, an opinion that, *ex hypothesi*, one arrived at in ignorance of the argument?[11]

However, in such cases much depends on the particular role that one's prior disbelief plays in leading one to conclude that the argument is not a good one. Thus, perhaps before encountering Socrates, Cebes is confident that *human beings do not possess immortal souls*. Attempting to convince him otherwise, Socrates offers an argument for the contrary conclusion. Contrast two cases:

> **Case 1**. Cebes reasons as follows: 'The conclusion of Socrates' argument is that human beings have immortal souls. But that's false. Because there are no sound arguments for false conclusions, the argument must harbor some hidden flaw.' He thus remains convinced that human beings do not have immortal souls.

> **Case 2**. Because Cebes is convinced that human beings do not have immortal souls, he believes that Socrates' argument for the contrary conclusion must harbor some hidden

[10] See, e.g., Jonathan St.B.T. Evans, J.L. Barston and Paul Pollard, 'On the Conflict Between Logic and Belief in Syllogistic Reasoning' *Memory and Cognition* 11(1983):295-306, and Evans, *Bias in Human Reasoning* (Lawrence Erlbaum: Hove, UK 1989).

[11] Indeed, attempts to characterize the distinction between good and bad arguments sometimes seek to do so partially in terms of the effects that such arguments would or wouldn't have upon idealized audiences of agnostics, neutral parties who neither affirm nor deny the conclusions of the arguments in advance. See, for example, Peter van Inwagen, *The Problem of Evil* (Oxford: Oxford University Press, 2006), Lecture 3, 'Philosophical Failure'.

> flaw. Because he believes that the argument harbors some hidden flaw, he scrutinizes it more thoroughly than he would have otherwise; because he scrutinizes it so thoroughly, he ultimately detects a subtle equivocation in the argument that would have escaped notice if he had subjected the argument to a level of scrutiny any less severe. He thus remains convinced that human beings do not have immortal souls.[12]

In each case, Cebes remains unmoved by Socrates' argument. Moreover, in each case, his doing so is counterfactually dependent on his prior conviction. However, the two cases differ crucially with respect to the relationship between Cebes' prior conviction and his reason for concluding that Socrates' argument is unsound. In Case 1, if asked what reason he has for thinking that Socrates' argument is unsound, Cebes will cite the fact that he believes that human beings do not have immortal souls, or (more likely) the proposition that *human beings do not have immortal souls* itself. This is the way of the Kripkean dogmatist. As we have emphasized, such reasoning is not generally legitimate.[13] In contrast, in Case 2, if Cebes is asked to defend his rejection of Socrates' argument as unsound, he will cite as his reason not the proposition that *human beings do not possess immortal souls*, nor the fact that he believes this proposition, but rather the flaw in Socrates' argument. In the second case, although Cebes' prior belief plays a crucial historical role in his recognition that Socrates' argument is unsound, the role in question is an essentially heuristic one: it belongs—to an invoke an old distinction—to the context of discovery, as opposed to the context of justification.

Notice that in the second case, unlike the first, Cebes' remaining unmoved in the face of Socrates' arguments is perfectly legitimate. After all, he has identified a genuine flaw

---

[12] Of course, neither the Cebes of Case 1 nor Case 2 bears much resemblance to Cebes as depicted by Plato in the *Phaedo*, who proves so deplorably acquiescent in the face of Socartes' sophistries.

[13] Plausibly, there are some cases in which the general pattern of reasoning *is* legitimate. For example, Zeno's contemporaries were, I think, justified in concluding that his arguments for the impossibility of motion are flawed, even if (as is almost surely the case) they lacked the philosophical and mathematical sophistication to say what is wrong with those arguments (i.e., what is wrong with them other than that their conclusion is false). In some cases, it might be more reasonable for one to think that the fact that an argument is flawless as far as one can tell is explained by one's ignorance or lack of sophistication rather than by its being flawless in fact. In general, the idea that there are 'Moorean facts'—roughly, propositions which we should treat as sufficient reasons to reject as unsound any philosophical argument which seeks to cast doubt on them—is one that enjoys considerable currency within contemporary philosophy. For references and further discussion, see my 'Moorean Facts and Belief Revision, or Can the Skeptic Win?' in John Hawthorne (ed.) *Philosophical Perspectives, vol.19: Epistemology* (Blackwell Publishers, 2005), pp.179-209.

in the argument. That he would not have done so were it not for the fact that he was antecedently disposed to believe that any argument for the conclusion in question must harbor some hidden flaw is of no normative significance. Indeed, his basis for declaring the argument unsound is no less strong than if the same reason were offered by someone who believed the argument's conclusion.[14]

What holds for formal fallacies in arguments holds for methodological problems in statistical studies as well: in the psychological studies which demonstrated the polarization phenomenon, individuals manifested heightened sensitivity to methodological problems in studies when the results of those studies seemed to tell against their beliefs. Indeed, psychologists who have discussed the phenomenon sometimes emphasize the extent to which individuals prove adept in identifying genuine limitations or weaknesses in studies that conflict with their prior beliefs.[15]

Of course, all of this might lead one to think that You and and I are guilty, not of giving too much scrutiny to evidence that seems to tell against our beliefs, but rather of giving too little scrutiny to evidence that seems to tell in their favor. Or better: perhaps our fault lies in the fact that we subject such evidence to *different* levels of scrutiny. That

---

[14] Compare a case which might at first glance seem even more suspicious, viz. a case in which one's *desire* not to believe the conclusion of an argument plays an essential role in the process which leads one to judge that that argument is unsound. In the Introduction to his *Philosophical Explanations* (Cambridge, MA: Belnap, 1981), Robert Nozick signals his intention to uncover flaws in formidable arguments that purport to show that we lack knowledge or free will, a project that (he explicitly declares) is motivated by the value to him of being able to conclude that we *do* possess knowledge and free will. (Presumably, equally formidable arguments the conclusions of which concern matters of relative indifference would be left unmolested.) Of course, if Nozick's desire to conclude that a given skeptical argument is unsound leads him to overestimate the force of his critique, then his remaining unmoved in the face of the argument is normatively inappropriate. However, if the desire leads him to uncover what is in fact a genuine flaw in the argument, then his remaining unmoved is perfectly legitimate. Indeed, in that case his basis for rejecting the argument is just as strong as if the same problem had been discovered by someone who passionately wanted skepticism to be true, and thus reported the relevant discovery in a spirit of great disappointment.

[15] See, e.g., Gilovich *op.cit.* p.54:

> The results of this experiment were striking. The participants considered the study that provided evidence consistent with their prior beliefs…to be a well-conducted piece of research that provided important evidence concerning the effectiveness of capital punishment. In contrast, they uncovered numerous flaws in the research that contradicted their initial beliefs…Now consider what the participants in this experiment did *not* do. They did not misconstrue the evidence against their position as more favorable than it really was. They correctly saw hostile findings as hostile findings. Nor did the participants simply ignore or dismiss these negative results. Instead, they carefully scrutinized the studies that produced these unwanted and unexpected findings and came up with criticisms that were largely appropriate…

is, perhaps whatever absolute level of scrutiny we ought to devote to newly encountered evidence—indeed, even if no absolute level of scrutiny is rationally required of us—in any case, the one thing that we are rationally required not to do is to devote different levels of scrutiny to evidence depending on how well it coheres with our prior beliefs. (Here as elsewhere, formal normative requirements, i.e., ones requiring consistency in some broad sense of that term, might seem easier to defend than more substantive ones.)

I will take up this natural thought shortly. First, however, I want to examine another psychological mechanism that seems to play a role in underwriting the polarization phenomenon, a mechanism that is structurally similar to the one we have just considered although somewhat more subtle in its operation. In fact, it is another manifestation of our tendency to devote more thought to evidence which seems to tell against our beliefs than to evidence which seems to tell in their favor:

> For a given body of data and a given hypothesis which purports to explain that data, the extent to which one is disposed to search for alternative explanations of the data is not independent of one's prior attitude toward the hypothesis.

Thus, suppose that one is presented with evidence E and that hypothesis H is a *potential explanation* of E: roughly, H is the sort of thing which, if true, would account for why E is true.[16] If one is convinced that H is true prior to learning that E is true, then, all else being equal, upon learning E one is disposed to treat H as the actual explanation of E and to increase one's confidence that H is true on the basis of E, which one treats as confirming evidence for H (at least, provided that one does not also already believe some alternative hypothesis which is also a potential explanation of E). If, on the other hand, one is convinced that H is *false* prior to learning E, then, upon learning E, one is more likely to search for some alternative explanation H' to account for E. And, all else being equal, the more cognitive resources one devotes to the task of searching for alternative explanations, the more likely one is hit upon such an explanation, if in fact there is an alternative to be found.

---

[16] This is somewhat overly simple as a characterization of what it is to be a potential explanation, but the complexities need not concern us here. For further discussion, see Peter Lipton, *Inference to the Best Explanation* (Routledge: London, 1991), especially chapter 4.

To illustrate with reference to the example of capital punishment: suppose that You and I are informed that two neighboring states, A and B, differ in that

Fact 1: State A, but not State B, has capital punishment, and
Fact 2: State A has a lower murder rate than State B.

The hypothesis of DETERRENCE is a potential explanation of Fact 2: it is the kind of hypothesis which, if true, would account for why Fact 2 holds. Given that I initially believe DETERRENCE, when I subsequently learn Fact 2, I am disposed to conclude straightaway that DETERRENCE is the actual explanation of that fact and to increase my confidence that DETERRENCE is true as a result. On the other hand, given that You initially disbelieve DETERRENCE, You are more likely to search for some alternative explanation in order to explain why Fact 2 holds. Suppose that as a result of your efforts, You do find some plausible alternative potential explanation. Having done so, You will increase the credence that you give to DETERRENCE in the light of Fact 2 to a lesser degree, inasmuch as You are aware of a plausible alternative potential explanation of which I am unaware. (One way of thinking about what is happening: for You but not for Me, the plausible alternative steals some of the credence that would otherwise go to DETERRENCE.) As You and I continue to respond to incoming evidence in the light of our prior beliefs in this way, the net effect is that we are pushed further and further apart. This then, is another aspect of a psychological model that is itself a rival hypothesis to Kripkean dogmatism as the mechanism which underwrites the phenomenon of polarization.

Let's suppose that this is in fact an accurate description of how our prior beliefs sometimes influence hypothesis generation. What normative significance (if any) would this have? Again, the normative issues that arise here are not completely straightforward. Of course, if one's conviction that some hypothesis is false leads one to try to explain away apparently supporting data by attributing them to some implausible and *ad hoc* hypothesis, then one's doing so is unjustified. On the other hand, suppose that one's conviction and the search that it prompts leads one to hit upon what is in fact a

formidable alternative explanation of the data, a hypothesis which does warrant serious consideration. The key epistemological fact here is the following:

> **The Key Epistemological Fact**: For a given body of evidence and a given hypothesis that purports to explain that evidence, how confident one should be that the hypothesis is true on the basis of the evidence depends on the space of alternative hypotheses of which one is aware.

In general, how strongly a given body of evidence confirms a hypothesis is not solely a matter of the intrinsic character of the evidence and the hypothesis. (Nor is it solely a matter of their intrinsic characters together with one's background theory of how the world works.) Rather, it also depends on the presence or absence of plausible competitors in the field. It is because of this that the mere articulation of a plausible alternative hypothesis can dramatically reduce how likely the original hypothesis is on one's present evidence.[17]

Consider an historical example that is often thought to illustrate this normative phenomenon. Many organisms manifest special characteristics that enable them to flourish in their typical environments. According to the **Design Hypothesis**, this is due to the fact that such organisms were so designed by an Intelligent Creator (i.e., God). The Design Hypothesis is a potential explanation of the relevant facts: if true, it would account for the facts in question. How well-supported is the Design Hypothesis by the relevant evidence? Plausibly, the introduction of the Darwinian Hypothesis as a competitor in the nineteenth century significantly diminished the support enjoyed by the Design Hypothesis. That is, even if there had been no reason to *prefer* the Darwinian Hypothesis to the Design Hypothesis, the mere fact that the Design Hypothesis was no longer the only potential explanation in the field tends to erode (to some extent at least)

---

[17] The point was forcefully pressed by Hilary Putnam in the 1960s as a reason for doubting that Carnap's vision for inductive logic was a well-conceived research program. The relevant papers are collected in his *Mathematics, Matter, and Method* (Cambridge: Cambridge University Press, 1975). Horwich *Probability and Evidence* (Cambridge: Cambridge University Press, 1982) concedes the general epistemological point but argues that a broadly Carnapian confirmation theory can successfully accommodate it. Chihara, 'Some Problems for Bayesian Confirmation Theory', *British Journal for the Philosophy of Science*, vol.38, pp.551-560 argues that orthodox Bayesianism founders on the same point. A good discussion of the general issue is Earman, *Bayes or Bust* (Cambridge, MA: MIT Press, 1992), chapter 7.

how much credence the Design Hypothesis merits on the basis of the relevant considerations.

For the sake of explicitness, let's bring the psychological and normative considerations together. As a psychological matter, when we encounter data that seem to go against what we believe, we are disposed to devote resources to the project of generating rival hypotheses to account for that data. To the extent that we are successful in generating plausible rivals, apparent counterevidence gets considered against a relatively rich space of alternative explanatory hypotheses. This fact tends to diminish the extent to which any particular hypothesis in the field gets confirmed or disconfirmed by the original evidence, inasmuch as the competitors tend to divide up the support conferred by the novel evidence among them. That is, the support which any one of the hypotheses receives is diluted by the presence of the others. (This last fact is a normative consequence of the operation of the relevant psychological process.) On the other hand, when we encounter evidence that is plausibly explained by things that we *already* believe, we typically do *not* devote additional resources attempting to generate alternatives. Data that seem to support hypotheses that are already believed thus tend to get considered against a comparatively impoverished or sparse background of alternative hypotheses. As a result of the less competitive milieu, the support conferred by the new evidence is not siphoned away, and thus tends to go in relatively undiluted form to the already accepted hypothesis. Over time, this invisible hand process tends to bestow a certain competitive advantage to our prior beliefs with respect to confirmation and disconfirmation.[18]

How do the psychological tendencies considered in this section—tendencies which apparently underwrite the polarization phenomenon--compare to Kripke-style dogmatism? In one important and salient respect, the way that You and I respond to new evidence resembles Kripke-style dogmatism, in that both essentially involve treating

[18] It proves a surprisingly delicate matter to give an account of the circumstances in which two potential explanations constitute *rival* hypotheses, i.e., when the explanations genuinely compete with one another for evidential support (as opposed to, say, supplementing one another as parts of some larger, more encompassing potential explanation). On this, see Harman, 'Competition for evidential support,' *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society* (Hillsdale, New Jersey; Lawrence Erlbaum: 1989): pp. 220-226.. Nevertheless, I assume that we have some intuitive grip on the notion of competition among explanatory hypotheses and can recognize particular instances of it, even if an abstract characterization proves elusive.

incoming evidence differently depending on how such evidence fits or fails to fit with one's prior beliefs. There is an important asymmetry in the way that we respond to evidence that seems to tell against our prior beliefs and evidence that seems to tell in favor of those beliefs. In another respect, the way You and I respond to our evidence differs significantly from Kripke-style dogmatism, in that evidence which seems to tell against our prior beliefs typically occasions *more* thought than other evidence (as opposed to simply being dismissed as misleading).

Still, aren't You and I every bit as unreasonable as the Kripkean dogmatist? The time has come to pursue this question in more depth.

4. Normative Implications

We should, I think, distinguish carefully between (i) questions about the rationality of devoting greater scrutiny to apparent counterevidence in the relevant ways and (ii) questions about the rationality or epistemic status of the beliefs that result from having done so. One might think that such circumspection is unnecessary. After all, it's natural to think that if the practice of devoting greater scrutiny to apparent counterevidence is unreasonable, then the beliefs at which one arrives by engaging in that practice are themselves unreasonable, and that, on the other hand, if the practice is not unreasonable, then the beliefs at which one arrives by engaging in the practice are not necessarily unreasonable either. However, there are good reasons to proceed cautiously here. On what I take to be the correct view of these matters, questions about (e.g.) how much time or effort one should devote to scrutinizing a given argument or piece of evidence are *practical* questions. Thus, whether it's reasonable for one to spend additional time pondering a given argument, or attempting to think of some alternative explanation of a given fact, might very well depend upon whether one has to leave immediately in order to catch one's flight. Typically, the reasons that one has to devote further thought to a given argument or piece of evidence (if any) compete with other practical considerations. In such cases, rationality is always in part a matter of opportunity cost, in the economists' sense. On the other hand, how confident it is reasonable for one to be that some proposition is true typically does *not* depend on considerations such as whether one has to

leave immediately in order to catch one's flight. Rather, how confident it is reasonable for one to be that some proposition is true is a matter (at least in paradigmatic cases) of how well-supported that proposition is by one's evidence.[19] In what follows, my ultimate focus will be on the epistemic status of the beliefs at which You and I arrive when we devote more thought to apparent counterevidence in the characteristic ways described above. I begin, however, by making some observations about the practice itself.

In considering the tendency to devote greater scrutiny to apparent counterevidence, we might picture someone who deliberately and self-consciously adopts this as a policy, perhaps with an eye towards maintaining or further bolstering his or her original views. (One resolves that one will devote more time and effort to searching for alternative explanations of data that seem to support hypotheses that one presently disbelieves, and so on.) However, it would be misleading, I think, to picture the characteristic tendency of individuals to devote more thought to counterevidence on this model, as the manifestation of a consciously adopted policy. On the contrary, the tendency to devote more thought to that which seems to violate or run counter to one's expectations would seem to be the natural or default state, which prevails unless one deliberately makes a conscious effort to devote equal thought to those considerations which seem to support what one already believes. If it is indeed unreasonable to devote greater scrutiny to phenomena that seem to violate one's prior beliefs or upset one's expectations, then this particular cognitive defect is a deeply-rooted one. Indeed, such lack of even-handedness would perhaps have some claim to being considered the Original Sin of Cognition.[20]

As we've seen, one manifestation of our lack of even-handedness in responding to new evidence is our tendency to devote fewer cognitive resources to searching for alternative explanations of a given fact when we already believe some hypothesis that

---

[19] For development and defense of these ideas, including further reflection on the relevant contrast, see my 'The Rationality of Belief and Some Other Propositional Attitudes' in *Philosophical Studies* 110 (2002): 163-196, and 'Epistemic Rationality as Instrumental Rationality: A Critique', in *Philosophy and Phenomenological Research* vol.LXVI, No.3 (2003): 612-640.

[20] Thus, consider one of the ways in which cognitive scientists attribute beliefs to pre-linguistic children. A sequence of events is produced within the child's visual field; the duration of the child's gaze is carefully timed. If the child continues to stare, this is taken as evidence that what has happened has violated the child's expectations, and as grounds for attributing the corresponding beliefs to him or her. On the other hand, if the child does not stare, no such attribution is made.

would account for that fact than when we do not. Isn't this just *obviously* an unreasonable practice? Here is a reason for thinking that it is not. Compare the practice of science. It is often claimed that the sciences (at least, the mature sciences) are to some extent *anomaly-driven*, in the following sense.[21] At any given time, there is a substantial range of phenomena that is well-accounted for by currently-accepted theory. The phenomena are exactly what one would expect given the truth of the accepted theory, the theory offers plausible and generally satisfying explanations of why particular events occur as they do, and so on. At the same time, there are various anomalies: salient phenomena that are not explicable in terms of the accepted theory, or worse, which stand in at least some *prima facie* tension with it. To this extent then, the anomalies seem to disconfirm or tell against the accepted theory.

Scientists do not treat the anomalous phenomena and the non-anomalous phenomena on a par. On the one hand, scientists devote relatively little attention and effort to attempting to devise plausible alternative explanations of phenomena for which the accepted theory *already* offers a plausible explanation. On the other hand, scientists devote a great deal of attention and effort attempting to generate hypotheses that allow the existence of the anomalies to be reconciled with the presently accepted theory (to the extent that such is possible). Assuming that this is in fact a fair characterization of one aspect of actual scientific practice, we can ask: are scientists unreasonable for behaving in this way? To what extent (if any) does their proceeding in this way impugn the rationality of science itself?

I don't believe that scientists are unreasonable for devoting more resources (intellectual or otherwise) attempting to generate novel explanations for anomalous phenomena than they do for phenomena that are already explained by the theory that they currently accept. (Indeed, one might very well think that to proceed in any *other* way would be unreasonable.) If this is correct, then the next question would seem to be the following: why think that what is reasonable in the context of scientific inquiry is unreasonable at the level of the individual thinker? Perhaps there is some reason for pulling the two apart. For example, perhaps any theory which is an accepted part of some

[21]The point is an especially prominent theme in Thomas Kuhn's *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press,1962) although it is, I believe, much more generally acknowledged than some of the more contentious claims of that work.

mature science has a kind of epistemic standing that is not enjoyed by the world view of any individual, and so it is incumbent upon any individual thinker to be even-handed in his treatment of new evidence in a way that it is not incumbent upon scientists *qua* scientists. Still, one might be understandably hesitant to declare that a practice which seems perfectly reasonable for scientists *qua* scientists (and indeed, which might very well be extremely conducive to scientific progress) is unreasonable when employed by ordinary thinkers. At the very least, there is a substantive issue here.

In general, the psychology of recent decades has not often shied away from claiming that human beings are significantly less rational than had traditionally been thought. It is noteworthy then, that some of the psychologists who have studied the polarization phenomenon have been quite reluctant to simply declare the relevant cognitive behavior unreasonable. Thus, Gilovich cautions as follows:

> At first blush, such uneven treatment of new information strikes most people as completely unjustified and potentially pernicious. It conjures up images, for example, of close-minded people disregarding a person's individual characteristics in deference to some invalid ethnic, gender, or occupational stereotype; it brings to mind examples of individuals and groups adhering to outmoded dogma…On closer inspection, however, the question of how impartial we should be in evaluating information that confirms or refutes our preconceptions is far more subtle and complicated than most people realize…the issue is complicated because it is also inappropriate and misguided to go through life weighing all facts equally and reconsidering one's beliefs anew each time an antagonistic fact is encountered…(*op.cit.*,pp.50-51).

Especially notable here is the view of Lord, Ross, and Lepper, the authors of the original capital punishment study:

> It is worth commenting explicitly about the normative status of our subjects' apparent biases…[T]here can be no real quarrel with a willingness to infer that studies supporting one's theory-based expectations are more probative than, or methodologically superior to, studies that contradict one's expectations. When an 'objective truth' is known or strongly assumed, then studies whose outcomes reflect that truth may reasonably be given greater credence than studies whose outcomes fail to reflect that truth (p.2106).

Indeed, Lord, Ross, and Lepper suggest that You and I are properly subject to criticism only insofar as our *initial* convictions are held more strongly than is warranted by our original evidence (pp.2106-2107).

Notice that this normative view, viz. that it is appropriate to give more weight to studies that cohere with one's prior opinions provided that those opinions are adequately justified by one's original evidence, would seem to license a modest form of Kripkean dogmatism. (Unlike full-fledge Kripkean dogmatism, one is not entitled to give zero weight to counterevidence, but one is permitted to discount such evidence to some significant extent.) Like full-fledged Kripkean dogmatism, the normative view in question can be shown to be false by appeal to the Commutativity of Evidence Principle. For again, let E represent some collection of evidence which has the following properties:

(i) There is some proper subset of E, E*, such that: if E* represented one's total evidence with respect to H, then one would be justified in believing that H is true on its basis, and

(ii) There is another proper subset of E, E**, such that: if E** represented one's total evidence with respect to H, then one would be justified in believing that H is false on its basis.

Suppose that over time, one gradually accumulates evidence that bears on H, until one's total relevant evidence consists of E. Now compare two different possible histories of how one might have arrived at that point. In Case 1, one first accumulates the evidence which comprises E*, forms the justified belief that H is true, and thus gives less weight to all of the subsequently considered evidence that counts against H. In Case 2, one first accumulates the evidence which comprises E**, forms the justified belief that H is false, and thus gives less weight to all of the subsequently considered evidence that counts in favor of H. One thus ends up more confident that H is true in Case 1 than in Case 2, despite the fact that one has the same evidence in both cases, because of purely historical facts about the temporal order in which the elements of E were acquired.

Indeed, one might think that the Commutativity of Evidence Principle can do even more work here. We have emphasized the significant differences between the way in which You and I respond to evidence in the light of our prior beliefs and the way in which the Kripkean dogmatist does. Still, one might think that the views at which You and I arrive by responding to our evidence in this way can also be shown to be unreasonable by appeal to the Commutativity of Evidence Principle, in a parallel manner. For if the 'alternative model' outlined in section 3 is in fact accurate as a descriptive

account, then it looks as though purely historical facts about the order in which You and I acquire our evidence will often make a difference to what we end up believing. Thus, consider some early time t0, before I had any opinion at all about whether the hypothesis of DETERRENCE is true. At time t1, I receive evidence which suggests that DETERRENCE is true, and I take up the corresponding belief in response. I thus respond to subsequently encountered evidence in the manner characteristic of a believer in DETERRENCE, as opposed to the manner characteristic of someone who disbelieves DETERRENCE or the manner characteristic of someone who neither believes nor disbelieves DETERRENCE. Thus, when at some still later time t2 I encounter evidence which seems to tell against DETERRENCE, I'm disposed to respond differently to this evidence in virtue of my belief. Among other things, I'm disposed to generate for consideration alternative explanations of the apparent counterevidence. I am thus more likely to consider the bearing of the apparent counterevidence to DETERRENCE against a relatively enriched background of alternatives, and this in turn tends to diminish the extent to which the apparent counterevidence disconfirms DETERRENCE for me, inasmuch as its probative force is somewhat diluted by the presence of the various rivals. On the other hand, suppose that I had encountered the same two pieces of evidence in the reverse order. If I had first encountered the evidence that suggests that DETERRENCE is false, then I would have taken up *that* belief, and I would respond to subsequently encountered evidence in the manner characteristic of someone who holds *it*. In particular, when I subsequently encounter the evidence that seems to suggest that DETERRENCE is true, it is this piece of evidence which is more likely to get considered against an enriched background of competing hypotheses and thereby to have its bearing on any one hypothesis lessened.

The suspiciousness of this is perhaps even greater when we focus once again on the interpersonal case of two individuals who have been exposed to both pieces of evidence, differing only in the order in which they encountered that evidence. If they both reason in the way described, the model predicts that they might very well end up with different levels of confidence towards the proposition that capital punishment is a deterrent, despite apparently having the same total evidence. In that case, it looks as though two individuals who share the same total evidence end up believing different things because

of historical facts about the relative order in which they encountered the elements that comprise that total evidence. If we say that each might nonetheless be reasonable in believing as he does, then this would seem to be a straightforward violation of the Commutativity of Evidence Principle.

However, this line of reasoning is mistaken. Initial appearances to the contrary, individuals in the above scenario do *not* violate the Principle when it is properly understood. This is because, when individuals reason in the envisaged way, they do not in fact end up with the same total evidence in the relevant sense. Here it's important to distinguish between two different senses of 'evidence', a broad sense and a narrow sense. Evidence in the **narrow** sense consists of relevant information about the world. Statistical information about crime rates is, perhaps, a paradigm of evidence in the narrow sense. As a rough rule of thumb: evidence in the narrow sense consists of things that it would be natural to call 'data'. In the narrow sense of evidence, the individuals in the scenario described above have the same total evidence. On the other hand, we can also speak about evidence in the **broad** sense. Evidence in the broad sense includes everything of which one is aware that makes a difference to what one is justified in believing. Clearly, evidence in the broad sense includes evidence in the narrow sense, inasmuch as relevant data or information of which one is aware typically does make a difference to what one is justified believing. But one's evidence in the broad sense will include, not only evidence in the narrow sense of data, but also things such as the space of alternative hypotheses of which one is aware. For (by the Key Epistemic Fact) which hypotheses one is aware of can make a difference to what one is justified in believing. Now, even if two individuals have exactly the same evidence in the narrow sense, they might have different evidence in broad sense, in virtue of differing with respect to the set of hypotheses of which they are aware. But if they have different evidence in the broad sense, then they might differ in what they are justified in believing, despite having exactly the same evidence in the sense of data. (Again, this will be admitted by anyone who accepts the Key Epistemic Fact.)

On the present view then, the following is true. For any given body of total evidence—where total evidence is understood as evidence in the broad sense-- the order in which the constituent pieces of evidence are acquired makes no difference to what it is

reasonable to believe. If one had arrived at the same body of total evidence by encountering the constituent pieces of evidence in a different order, one would be justified in believing exactly what one is justified in believing as things actually stand. Thus, the Commutativity of Evidence Principle is respected. On the other hand, historical facts about when one acquires a given piece of evidence might make a *causal* difference to which body of total evidence one ultimately ends up with. One acquires a given piece of evidence at an early stage of inquiry; this might very well influence the subsequent course of inquiry in various ways, by way of making a difference to how one subsequently thinks and acts (which possibilities one considers, which routes get explored as the most promising and fruitful, and so on.) And this in turn can make a difference to what evidence one ends up with. In such cases, there is an undeniable element of path-dependence. It is an interesting question, I think, how troubled we should be by the specter of such path-dependence (if we should be troubled at all). Is it enough to undermine the reasonableness of one's believing as one does, that one might very easily have arrived at a different body of total evidence, that one's having arrived at *this* particular body of evidence is in various ways a highly contingent, fragile matter? (In some extremely close possible worlds, one's total evidence is significantly different.) I'm not convinced that it is: I think that if one's beliefs are ones that it would otherwise be reasonable to hold in the light of one's total evidence, then the fact that it is a highly contingent, fragile matter that one has this particular body of total evidence rather than some other is not enough to undermine the reasonableness of one's believing as one does.

However, one might think that there is a special feature with respect to the case at hand. Here, not only does one know that one easily could have had different total evidence, but one also has some idea about the *direction* in which one's actual total evidence is likely to be skewed, viz. it is likely to be skewed in the direction of those beliefs that one held at the outset. One might then think that one ought to correct for the operation of the relevant psychological mechanisms, by being less confident of those beliefs that are likely to have been the past beneficiaries of the mechanisms. In short, to the extent that the invisible hand becomes visible, one ought to correct for its operation.

I believe that this last thought is correct. Those few of us who are *aware* of the phenomenon of belief polarization—which includes, presumably, attentive readers of the

present paper-- ought to be less confident of beliefs that are likely to have benefited from the underlying psychological mechanisms. The psychological mechanisms in question constitute *biasing factors* inasmuch as they influence the evidence which one ends up with in a systematic, directed way. (That is, the evidence one ends up with is likely to be a biased sample of the evidence that one would have had if the relevant psychological mechanisms were not operative.)

From this, of course, it doesn't follow that the average person who is presumably unaware of the phenomenon of belief polarization is unreasonable in believing in accordance with his or her total evidence--even if her having that body of total evidence rather than some other is partially due to the past operation of the relevant kind of biasing factors. In general, the fact that distorting or biasing factors played a role in one's arriving at total evidence E does not make it unreasonable to believe in accordance with E, provided that one is unaware of the operation of those factors; what would be unreasonable would be to fail to adjust one's views upon learning of the role played by those distorting or biasing factors. Thus, suppose that you are my only source of information about what kind of person Leopold is, and I have no reason to distrust your reports on the subject. Nevertheless, you always pass along any information about Leopold that casts him in an unfavorable light while systematically withholding information that casts him in a favorable one. In these circumstances, it is not unreasonable for me to hold a negative opinion of Leopold on the basis of the information available to me; what would be unreasonable would be to fail to adjust my view upon learning of your role in biasing my evidence with respect to the question. (In this latter case, of course, my total evidence would have changed in a crucial way.)

In general, accurately proportioning one's beliefs to one's total evidence suffices for believing reasonably. But facts of which one is completely unaware are not eligible for inclusion among one's total evidence. For this reason, I think that we should admit that the beliefs of someone who responds to evidence in the way described here can be reasonable, provided that he is completely unaware of the fact that his evidence is likely to be biased in this way. In presenting these ideas in various forums, I have found considerable sympathy for this verdict, but also some resistance, as well as no small amount of ambivalence. I will end this section by offering a speculative diagnosis of why

many of us--for I include myself here--tend to have somewhat soft intuitions about this sort of case. When you pass along information that casts Leopold in an unfavorable light, while filtering out information that casts him in a favorable one, the evidence which I end up with is in effect a biased sample of the evidence that I would have had, had you not acted in this way (and no similar distorting factor had operated instead). That my belief is nonetheless a reasonable one, despite being based on an unrepresentative sample of evidence, is due to the fact that I am nonculpably oblivious to this. But perhaps there is also another factor that is relevant here: the biasing factor is completely *external* to me, not only in the sense that it operates wholly outside of my ken, but also in the sense that my own agency plays no role in the relevant mechanism. Notice that in this respect, a person who subjects apparent counterevidence to greater scrutiny (and thus tends to arrive at what is in fact a biased sample of the evidence that he would have wound up with otherwise) but is non-culpably ignorant of this, seems to constitute something of an intermediate case. On the one hand, he is unaware of the fact that a biasing factor played a role in his arriving at this body of total evidence. On the other hand, his agency is complicit in the fact that he now possesses a biased sample of evidence; the biasing mechanism is located in *him*. Perhaps this accounts for why intuitions about the status of the beliefs arrived at in this way tend to be less firm than intuitions about more paradigmatic cases of rationality and irrationality.[22]

5. Conclusion

The following is, I believe, a not uncommon pattern. Relatively early on, one picks up a view about some controversial matter, a view that is not shared—and indeed, is explicitly rejected--by some who have considered the question. Perhaps one even picks up the view at One's Parent's Knee. Once one first begins to hold the view, one retains it thereafter. [23] Perhaps at various times one is somewhat more confident than at other

[22] The concept of bias is, I believe, one that could stand more analytical hatchet work than it has thus far received. For some early stabs, see Nozick, *The Nature of Rationality* (Princeton, NJ: Princeton University Press, 1993), pp.100-106, and my 'On Following the Argument Where It Leads' (in preparation).

[23] A recent, particularly interesting autobiographical account of this phenomenon by a philosopher is G.A. Cohen, *If You're An Egalitarian, How Come You're So Rich?* (Cambridge, MA: Harvard University Press,

times, but after one first comes to hold the view, one can from then on be correctly described as believing the relevant proposition. Over time, however, the reasons for which one holds the view evolve. That is, the reasons for which one believes that [EXAMPLES] are not identical to the reasons for which one held this belief, when one first began to hold it. (If pressed to defend one's view now, the considerations that one would cite are different from the considerations that one would have cited then.) Indeed, perhaps reflection on one's past self would prompt thoughts of the following sort:

> Looking back on it, the reasons for which I first came to hold this view were not particularly strong. Indeed, given the considerations available to me then, I was probably overly confident. However, this purely biographical fact is not relevant to how confident I should be that the same belief is true now. For how confident I should be now depends purely on the reasons for and against the belief that I currently possess. Thus, even if at some point in the past I was overly confident, this is no reason for me to be any less confident of the view now, for I currently have stronger reasons for thinking that the view is true than I did then, reasons which *do* suffice to justify my present level of confidence. For me to think that the quality of the reasons for which my past self held the belief is somehow relevant to what I should think now would be to commit a version of the Genetic fallacy.

This line of thought might seem unimpeachable. But for reasons that can perhaps be anticipated given the discussion to this point, I think that it proceeds too quickly. There are several reasons why some measure of suspicion seems in order in the circumstances.[24] The point that I wish to emphasize is the following. Even if one can reasonably assume that one is giving due weight to all of the relevant considerations of

---

2000). See especially Chapters 1 and 2, 'Paradoxes of Conviction', and 'A Montreal Communist Jewish Childhood'.

It would be interesting to know, although no doubt difficult to discover, how common it is for individuals to be aware of controversial issues for some significant length of time before first forming opinions about them. For my own part, I confess that I cannot remember a time when I was aware of issues such as the moral permissibility of abortion, or the moral permissibility of eating red meat, or whether human beings possess free will, but had not yet formed an opinion; as far as I can tell, my awareness of each of these issues—as well as countless others-- is more or less coeval with my having some opinion or other about them. (I do not report this fact with pride.)

[24] In addition to the reason cited in the main text, there is this: to the extent that one now judges that one's past reasons for holding the view in question were not sufficient to justify one's past attitude towards it, one gains some negative inductive evidence about how reliable one is in weighing evidence of the relevant sort (presumably, one is making a mistake at some point, either now or then). I do not want to press this point too hard, however. Among other things: particularly in a case in which one first formed the relevant belief relatively early on in life, perhaps one can reasonably assume that one's ability to accurately assess evidence of the relevant sort has improved with greater intellectual maturity.

which one is currently aware, there are still reasons for suspicion when the belief has the relevant kind of history. For as we have seen, the fact that a belief is held at earlier times can skew the total evidence that is available at later times, *via* characteristic biasing mechanisms, in a direction that is favorable to itself. The concern is not (simply) the banal point that an individual who has long held a given view might easily fall into overestimating how well-supported it is by the considerations available to him; rather, the very fact that he has this particular body of considerations available, rather than one that is significantly less favorable, might very well be due to the fact that he has long been a believer.[25] In deciding what level of confidence is appropriate, we should taken into account the tendency of beliefs to serve as agents in their own confirmation. Moreover, inasmuch as the possibility that the relevant biasing mechanisms played a role in skewing one's total evidence is a cause for concern even when one's original belief *was* initially based on adequate evidence, the reasons for concern would seem to be even stronger in a case in which one now judges that one's earlier reasons were not particularly strong. Thus, I'm inclined to think that, unless one has some special reason to think that one does not respond to apparent counterevidence in the way that individuals typically do, one should be less confident of beliefs with the relevant kind of history.

Descartes initiated modern philosophy when he embarked upon an intellectual project of immense ambition. According to his own account, the project was inspired by his doubts about a view of the world that had been built upon opinions uncritically inherited in his youth. Concerned that his attempts to achieve anything worthwhile in the sciences would inevitably be undermined by the influence of such opinions, he set out to begin anew, by suspending his commitment to everything that he had previously taken for granted. In attempting to determine what is true, he would begin with a cognitively clean slate. To do otherwise would be to load the dice at the very outset of inquiry, in a way that would risk tainting its subsequent deliverances.

By the twentieth century, if not earlier, this Cartesian project had become a popular target for detractors, including some thinkers of the highest rank. Peirce saw the

[25] In this respect then, it would seem to make sense for an individual so situated to 'distrust reason' to a certain extent. See the stimulating discussion in Hilary Kornblith's excellent essay 'Distrusting Reason', *Midwest Studies in Philosophy*, XIII (1999), pp.181-196.

Cartesian aspiration to begin from a cognitively clean slate as naïve at best and an invitation to stultifying pretense and self-deception at worst.[26] Quine, following Neurath, repeatedly counseled us to think of our cognitive situation as analogous to the plight of sailors attempting to repair their ship on the open sea. Although particular planks might be removed, such change is of necessity piecemeal in character. Similarly, although particular pieces of the web of belief might be replaced (typically peripheral ones), any such change takes place against the unquestioned background provided by the rest of the web; we can never stand outside the web of belief altogether. We never Start from Scratch.

Perhaps it is true that the kind of cognitive purity which Descartes sought at the outset of his own inquiry is not a state which we can reasonably hope to attain. Still, even if that's so, we ought not to be cavalier about this fact or to underestimate the potential costs which accompany it. And we ought not to despise the Cartesian aspiration to attain a kind of strong neutrality and objectivity, a position from which future inquiry might be conducted in such a way as to be maximally safe from being compromised by the seemingly inevitable weight of past opinion. For from the present vantage point, the radical nature of the Cartesian project seems indicative of its author's unusual sensitivity to what is in fact an all too pervasive phenomenon.[27]

---

[26] 'We must begin with all the prejudices which we actually have when we enter upon the study of philosophy. These prejudices are not to be dispelled by a maxim, for they are things which it does not occur to us can be questioned. Hence, this initial skepticism will be mere self-deception, and not real doubt…' (*Collected Papers of Charles Sanders Peirce*, edited by Charles Hartshorne, Paul Weiss, and A. Burks (Cambridge: Harvard University Press, 1931-1958), 2.265).

And elsewhere:

> 'There is but one state of mind from which you can 'set out', namely, the very state of mind in which you actually find yourself at the time you do 'set out'—a state in which you are laden with an immense mass of cognition already formed, of which you cannot divest yourself if you would…'(*op.cit.* 5.416)

[27]Ancestors of this paper were presented at Dartmouth College, the 2005 APA Central Division Meetings, (as my contribution to an invited symposium on the concept of 'Evidence') and at meetings of my graduate seminars at Princeton in the springs of 2005 and 2006; I am grateful to the audiences present on the occasions. In addition, I would like to thank Walter Sinnot-Armstrong, Dan Garber, Emily Pronin, Adam Elga, Joshua Knobe, Roy Sorensen, Mark Budolfson, Jose Luis Bermudez, Isaac Choi, and especially, Marian David, my respondent at the aforementioned 'evidence' symposium.