Synthese (2014) 191:2529–2547
DOI 10.1007/s11229-014-0438-x

False polarization: debiasing as applied social epistemology

Tim Kenyon

Received: 15 January 2014 / Accepted: 15 January 2014 / Published online: 20 March 2014

Abstract False polarization (FP) is an interpersonal bias on judgement, the effect of which is to lead people in contexts of disagreement to overestimate the differences between their respective views. I propose to treat FP as a problem of applied social epistemology—a barrier to reliable belief-formation in certain social domains—and to ask how best one may debias for FP. This inquiry leads more generally into questions about effective debiasing strategies; on this front, considerable empirical evidence suggests that intuitively attractive strategies for debiasing are not very effective, while more effective strategies are neither intuitive nor likely to be easily implemented. The supports for more effective debiasing seem either to be inherently social and cooperative, or at least to presuppose social efforts to create physical or decision-making infrastructure for mitigating bias. The upshot, I argue, is that becoming a less biased epistemic agent is a thoroughly socialized project.

Keywords Social epistemology · Bias · Debiasing · False polarization

1 False polarization

I recently attended a public lecture, the thesis of which was that diversity of gender and race among science workers enhances the epistemic virtues of the scientific enterprise. The speaker offered significant empirical evidence that diversity in the population of scientists doesn't just make for social justice; it makes for better science.

During the question period, it became clear that one questioner was clearly vexed by the talk. The questioner insisted, repeatedly and at length, that Newton and Darwin had been great scientists even though there was virtually no diversity of the relevant

T. Kenyon (✉)
University of Waterloo, Waterloo, Canada
e-mail: tkenyon@uwaterloo.ca

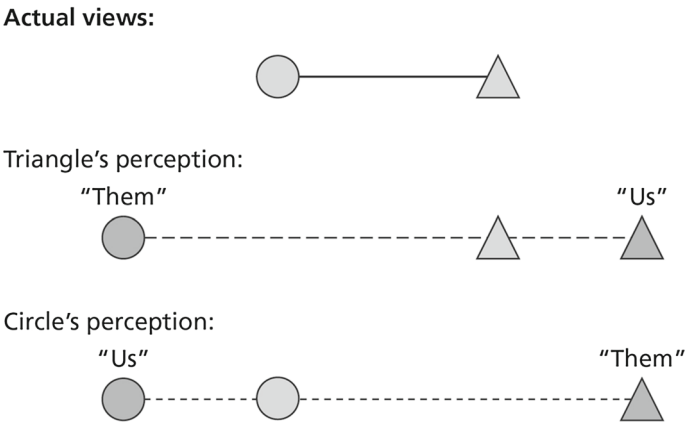

Fig. 1 Adapted from Pronin et al. (2002a)

sorts in the scientific communities they inhabited. It was clear, from the way this was presented as an objection, that the questioner understood the speaker to be committed to the view that, without diversity of race and gender, there could be *no* good science; or perhaps even that white Anglo-Saxon men could not be great scientists. Confronted with (what the questioner perceived as) such a radical and intellectually bankrupt speaker, the questioner was derisive and dismissive. Indeed, the questioner was so obviously hostile as to make following up the conversation an unappealing prospect for the speaker, who had neither defended nor even hinted at such extreme claims.

How did the questioner attribute to the speaker a complex of views that would have been falsified by the Newton and Darwin examples? Plausibly, the questioner perceived the speaker as defending elements of a broadly progressive or reformist position regarding science and gender. In defending *some* aspects of this position, to some degree, the speaker was interpreted as holding *all* its notional elements, to the most extreme degree. The upshot was the perception, on at least one side of the exchange, of a radical disagreement. Communication quickly broke down, and was not reestablished.

It is perhaps not much of an insight to note that fruitful communication can be hampered by highly polarized beliefs. As viewpoints become more radically opposed in a discursive context, the prospects worsen for finding common ground from which either disputant may be engaged in productive discussion. This much is practically a platitude. What is perhaps less immediately obvious is that fruitful communication is hampered even by a *false* perception of polarization. This is the scenario manifest in the *false polarization* effect (FP). This term denotes the tendency for disputants to overestimate the extent to which they disagree about whatever contested question is at hand (Keltner and Robinson 1993; Robinson et al. 1995; Keltner and Robinson 1996; Puccio and Ross 1998; Pronin et al. 2002a; Monin and Norton 2003). The result can be visualized by means of Fig. 1.

In this diagram, Circle is aware of the respects in which her own view is more moderate or nuanced than the most radical stereotype version of the circular view.*Mutatis mutandis*, the same applies to Triangle and the triangular view. Nevertheless, each

disputant assimilates her interlocutor's view to the opposing stereotypical position. Each thereby comes to overestimate the extent to which their actual views differ. So, for example, someone who believes that abortion is in some respects deeply morally problematic might still believe that abortion is morally permissible in the first two trimesters, or that abortion should be freely available to victims of rape. Because she knows these facts about her own views, she sees herself as moderate, relative to the most extreme anti-abortion positions that she believes to exist. Yet when she hears someone else say that abortion services should be widely available, she is inclined to interpret that person as holding that all abortion, at any stage of pregnancy, is completely morally unproblematic—she assimilates her interlocutor to the stereotype of the most extreme opposing position. And a similar effect may apply in the opposite direction.

In this way, debates between people who in fact agree on some non-trivial propositions—e.g., "Other things being equal, the fewer abortions the better" or "Late-term abortions are more ethically fraught than early-term abortions"—are sent off the rails by a common misunderstanding. Speakers may moreover see their interlocutors as biased, or as bearing some heavy burden of proof (Kennedy and Pronin 2008). Indeed, each may be inclined to take it as dishonesty or coyness that her interlocutor will not come right out and articulate the most extreme views she (notionally) holds. This too could exacerbate the distrust between them.

While practically any cues that signal the holding of a perspective could initiate false polarization reasoning, it stands to reason that imprecisely defined labels for positions can be especially important to facilitating FP in normal dialogue. This is illustrated by two more sample exchanges.

Circle: "I just think there's something really arrogant or hubristic about atheism."
Triangle: "Interesting. I'm an atheist myself."

Here, let's suppose, Circle takes atheism to be the view that it is known, with mathematical certainty, that the universe contains no gods. To be an atheist, Circle thinks, is therefore to have complete confidence that no region of the universe contains *anything* answering to *any* definition of 'god'. By contrast, when Triangle calls herself an atheist, she just means this to indicate that she believes on the whole that, on the balance of evidence, there are no gods. She means this in more or less the same way she believes there are no unicorns: not as a matter of certainty, but with great confidence, because there is no good reason to believe that there are any. (Circle would call this agnosticism.) In this case, we could suppose that Triangle and Circle actually agree that it is possible that a god exists. They agree that the best arguments against theism are (merely) evidential. They agree that it is possible to survey the available evidence in good faith and yet not positively accept theism. Nevertheless, the differences in the way they label the positions means that each regards the other as extremely unreasonable, following the exchange given above. Circle regards Triangle as arrogant and hubristic, while Triangle regards Circle as ignorant and grossly uncharitable.

Triangle: "I'm sick of free-speech fundamentalists who go on and on about the evils of political correctness."
Circle: "Well, for my part, I'm absolutely committed to freedom of expression."

This could be a nasty case of FP, even if Circle and Triangle unknowingly agree that freedom of expression should be legally protected in nearly every case, but that there are important exceptions. Both may agree that political criticism should be absolutely protected. They may agree that shouting "Fire!" in a crowded theatre should not be legally protected behaviour. Both may agree that, e.g., the pornographic depiction of sexual violence is a problematic case of free expression, even in the event that they happen to disagree about its legal protection. Circle's attitude might be: *Violent pornography is awful and probably harmful, but the best approach on balance is to hold your nose and tolerate it*. Triangle's view might be: *The most extreme violent pornography is not legally protected, but censoring it requires a strong rationale, periodically revisited via transparent debate, to outweigh the default presumption of legal protection for artistic expression*. Indeed, it could happen that Circle and Triangle agree on *every* tenet of legal and political importance on this issue, with Triangle simply focusing on the significance of the worst cases (perhaps due to having had frustrating discussions on the topic in the past), and Circle emphasizing the breadth and significance of the legally protected class of cases. Nevertheless, FP could leave them polarized and uncommunicative.

To the extent that we value the fruitful exchange of ideas, a worry is that partisans or mutually perceived opponents in a FP context

> are apt to underestimate the possibility of finding common ground that could provide the basis for conciliation and constructive action; as a consequence, they could be reluctant to enter into the type of frank dialogue that could reveal such commonalities in interests or beliefs (Robinson et al. 1995, p. 405).

If either side in such a dispute is subject to FP, the prospects for constructive dialogue may be reduced. If both sides are affected, then the effective difference between the two viewpoints will include an aggregate of the interlocutors' mutual misunderstandings, which can "lead the conflicting partisans to see the side as extreme, unreasonable, and unreachable" (Pronin et al. 2002a, p. 651). As a matter of applied social epistemology, it therefore seems desirable to have strategies for avoiding or minimizing FP. That is, we should seek ways of *debiasing* for false polarization.

In the following remarks I examine FP as an instance of the problem of teaching and learning the skills of forming reliable beliefs about other people. It is, in short, a problem of applied social epistemology. As a barrier to the formation of accurate beliefs about one's interlocutors, and as an impediment to the exchange of information, FP fairly clearly undermines what are sometimes called *reliable belief-forming processes*.[1] If possible, then, the cognitive tool-kit of an agent who cultivates such reliable mechanisms of doxastic fixation ought to contain the means to mitigate FP. But how does one acquire such skills? Can debiasing in general be learned and deployed as a self-administered treatment, and can debiasing for the false polarization of beliefs in particular be taught, learned, and used?

[1] The notion of a reliable belief-forming process plays a key conceptual role in some versions of epistemic reliabilism (Goldman 1986, pp. 44–51); but one need not be an epistemic reliabilist to think that reliable belief-forming processes are important things to cultivate.

Setting aside a very few scholarly expressions of pessimism (e.g., Willingham 2007), societies and educational systems the world over have largely voted with their wallets on this question: enormous resources are devoted to teaching the skills of reliable, undistorted reasoning all over the world.[2] More specifically, thousands of courses in critical reasoning, epistemology, cognitive psychology, and rhetoric are taught every year (some of them by me), with at least the partial aim of imparting to students the ability to significantly mitigate biased reasoning. But under what circumstances, and with what constraints, might this be a plausible aim? When we take seriously the best available empirical evidence, I think, we must acknowledge not just the difficulty of acquiring the skills to mitigate FP in particular, but the general problem of teaching and learning effective debiasing strategies. We will moreover be led to recognize the profoundly social and collective character of those belief-forming processes that effectively minimize distorting biases.

Naturally, there is a great deal more data and theorizing about mitigating biases in general than about mitigating FP in particular. But in some key respects, that literature converges on common conclusions, especially regarding the mitigation of biases. So I will begin by sketching some evidence regarding the general difficulty of debiasing, before considering the specific case of FP. The current state of evidence suggests a number of conclusions, which have not been widely appreciated in aggregate in social epistemology (or in educational theory, for that matter). First, merely teaching people *about* biases does not reliably lead to debiasing. It can even make people's judgements less accurate. Moreover, the debiasing strategies that do seem to work under experimental conditions for at least some important biases are relatively complicated, artificial, and, if not counterintuitive, are at least not especially intuitive. These strategies are apt to be quite difficult for people to implement individually, in informal circumstances. And while there is some reason to think that the task of mitigating FP will escape at least one widespread problem with debiasing—the problem of overcorrection, as I will argue—there are other problems linked to real and perceived social pressures that arise in particularly strong forms for the everyday debiasing of FP.

The upshot, I will argue, is that an atomistic or individualistic approach to mitigating FP is especially unlikely to succeed. There are reasons to think that successfully remedying socially complex biases will require a rich framework of educational background, supportive social attitudes and practices, and specialized institutional infrastructure. The practical epistemology of debiasing thus turns out to be social in a very rich sense: that is, only in the presence of significant socio-institutional scaffolding is there realistic hope of minimizing biases like FP in real-life situations. The projects of being an epistemically responsible believer, and of educating people to have more reliable, less distorted, belief-forming dispositions, are inherently socialized ones.

How broadly are these conclusions intended to apply? I have already alluded to some commonalities across different biases with respect to the mitigation of distortions on

[2] Willingham writes, "Can critical thinking actually be taught? Decades of cognitive research point to a disappointing answer: not really" (2007, p. 8). Yet his skepticism is mainly directed at the way critical thinking is *typically* taught—that is, as a small set of general purpose skills, and with little emphasis on the ties between critical reasoning and the content of the claims at issue. This is probably not a blanket rejection of the utility of teaching critical thinking.

reasoning. This might seem quite surprising in itself, given that the differences between kinds of biases plausibly reflect differences in the cognitive and affective mechanisms of bias implementation. Why, then, should there be much commonality across effective bias-reduction strategies? The most important response to this worry is to qualify the scope of the claims regarding debiasing strategies, *inter alia*, which I will indeed do throughout these remarks; with such a range of phenomena at issue it is incautious to do much more than suggest conclusions. It is also worth noting, however, that it would not be particularly surprising if a wide range of distinct complex social biases were aggregate effects of more primitive or basic phenomena, such as confirmation bias, the availability heuristic, or naïve realism.[3] This would in turn make the existence of a relatively small set of effective bias-mitigating strategies less surprising, since disrupting the operation of a constituent bias would potentially mitigate the distorting effects of the aggregate bias.

A final methodological note emphasizes two related decisions regarding the focus of discussion. One is to emphasize debiasing the *judgement*, rather than the *agent*, for reasons I canvass below regarding the ephemerality of debiasing effects, even when these are successful. Another is to emphasize practical reasoning and action in debiasing rather than just the cognitive element. That is, by focusing not just on an agent's long-term ways of thinking but on external infrastructure that modifies the agent's environment, we end up implicating all sides of the fuzzy boundaries between (i) strategies that affect the way agents act by affecting the way they think generally, (ii) those that affect action in context by promoting feedback and regulation of cognition in context, and (iii) those that simply constrain action, quite apart from the thinking that produced it, in ways that filter out foreseeable effects of biases. I will not lean on that tripartite distinction in what follows, but I submit that it is there to be found in the approach that broadly emphasizes the importance of socio-institutional infrastructure in the promotion of effective debiasing.

2 Debiasing is hard

A significant general problem with mitigating the distorting effects of biases is that we are apt to go about it on the basis of intuitions or naïve theories about learning and judgement that are misguided, for all that they are powerfully appealing. "Just as people are at the mercy of their theories when deciding whether a response is biased," note Wilson et al., "so are they at the mercy of their theories when deciding how to correct for this bias" (Wilson et al. 2002, p. 191). The conviction that errors and distortions arise from ignorance, for example, is hard to hold at arm's length. On this view, problematic biases would be mitigated merely by learning about them, and by learning about their propensities to influence the behavior. Similarly, one might find it pre-theoretically plausible that implicit reasoning can generally be made explicit through careful reflection, so that implicit distortions of judgement could be remedied

[3] Lilienfeld et al. (Lilienfeld et al. 2009, pp. 392–394) briefly discuss these relatively primitive biases, proposing both that they can be partially constitutive of larger-scale phenomena (such as ideological extremism), and that they function *individually* as potential barriers to mitigating the biases with which they are linked.

by reviewing one's reasoning explicitly. Convinced of such general diagnoses, though, one would expect that learning facts about biases and showing individualistic mental discipline would suffice to minimize biases like FP. This would be a mistake.[4]

Teaching people facts about biases is relatively easy; teaching them to apply those facts to themselves at the right time, in the right way, is often extraordinarily difficult. The interesting point for our purposes isn't simply that debiasing is hard. It is that debiasing is hard even when one has a robust grasp of several things that our naïve theories suggest should facilitate it. This includes understanding the nature of the bias; the circumstances under which it can be active; and the fact that one is oneself susceptible to it in principle. There are various reasons for this surprising immunity of many biases to mitigation by straightforward education about them. I will briefly consider three interrelated ones.

One explanation for this is our susceptibility to *bias blindspot* (Pronin et al. 2002b). This term refers to the tendency for people to notice bias in others' judgements and behavior, while disregarding the prospect that they themselves are, in the event, also subject to the bias in question. This is the cognitive version of a more general phenomenon we might call a *skill deficit blindspot*. To know that people in general are prone to a particular kind of failing under certain circumstances, and even that one is oneself therefore susceptible in principle to that kind of failure under those circumstances, is not the same as being disposed to recognize one's failure when those circumstances hold.[5] Manifesting a bias blindspot does not require that one take oneself to be immune from bias. I may very well allow that I am biased on many occasions while yet holding that *this* is not one of those occasions. The blindspot in this type of case is the tendency for me to judge that the occasion at hand is always one of the unbiased ones.

A second reason it is difficult to acquire or impart the skill of debiasing is that we easily take ourselves to have successfully debiased when we have not. This might itself be understood as a higher-order version of bias blindspot, or as a rationalization in support of the blindspot. Not only is this problem not *mitigated* by a theoretical knowledge of the need to adjust for bias—it can be *enabled* by that knowledge (Thompson 1995). Armed with a theoretical conception of some bias, and a sense that the situation at hand is one in which I might be subject to the bias, I have the raw materials to confabulate a justification of my actually biased judgements or perceptions. I can introspect on my motives, conduct a personal inventory of my evidence for my belief, and conclude that I am not (unduly) biased. However, both this conclusion and the way I go about checking it may nevertheless be manifestations of the very bias against which I thought I was proofing myself. The idea that one can debias merely by firmly

[4] In the same vein, it might seem natural to suppose that FP arises simply from the cognitive or emotional investment that discussants have in a topic. The evidence suggests, however, that FP is unlikely to be entirely a matter of the partisanship of interlocutors. Pronin, Puccio and Ross note two studies finding that even people who were nonpartisan with respect to some socially charged topic also tended to overestimate the degree of disagreement between partisan interlocutors (2002, p. 652). Partisanship just seems to make us worse at something we're independently not great at doing: accurately estimating the prospects for conciliation between people who disagree over some fraught issue.

[5] This much applies to many kinds of human activity. For example, Strayer et al. (2012) note that drivers can enthusiastically criticize the poor performance that other drivers display when using mobile phones behind the wheel, while describing their own driving-while-phoning as safe.

thinking it over, that debiasing can be a cognitive act of will, is intuitively tempting but mistaken (Frantz and Janoff-Bulman 2000).

One complicating factor for the accuracy of self-perceptions of bias is that the influence of a bias on an agent shifts over time. Grant that I succeed in debiasing my judgement on one occasion, and accurately perceive that I have done so; yet by resting on my laurels and pronouncing myself bias-free (in at least that one respect), I set the stage for subsequent biased judgements, made under the illusion of persisting freedom from distortion. The point is made lucidly by John Randolph, who uses an arresting metaphor to describe racist biases. Randolph observes that it is tempting to think of these biases as analogous to tonsils, when they are in fact analogous to dental plaque. What's the difference? Once tonsils are removed, they are gone for good; but plaque collects gradually in our teeth forever, and requires a constant program of dental hygiene to counteract it (Randolph 2012).[6] Not only are we apt to believe that we have successfully debiased when we have not, but we may over-interpret the significance of a genuine debiasing success as well—taking ourselves to be free and clear thereafter, rather than having experienced just a merely fleeting success. Either way, telling oneself that one has debiased is a very different thing from actually minimizing distorted thinking in the case at hand.

A third reason it is hard to teach or learn how to debias is an extension of both preceding points. Not only does self-debiasing often appear unnecessary from the perspective of the biased agent, and not only does it often fail when attempted, but it can actually strengthen or consolidate biases (Hirt and Markman 1995; Sanna et al. 2002). Mistakenly telling oneself that one has debiased can leave one feeling even more strongly justified in holding the biased belief, despite not having genuinely vetted it.

Consider a passage posted on a blog by a fourth-year engineering student at my home institution:

> I feel, in some very specific ways, that Waterloo is better than MIT/ Harvard/ Stanford. And this is even after de-biasing myself because I went to Waterloo and that I will have paid $80,000 in total for the entire process.[7]

The student shows wise caution in noting that having a large financial investment in a project is apt to strongly influence one's evaluation of it. But the author's amateur status as a debiaser (along with the reflexive construction 'debiasing myself') strongly suggests that the attempted debiasing strategy was just that of an introspective mental effort to factor out the financial aspect of the evaluation. The phrasing powerfully conveys the sense that the author is more confident in the judgement in light of his conviction that the judgement was vetted by a self-debiasing process. Set aside the question of whether the judgement is correct; still, plausibly, we should regard this as a case of belief becoming more strongly held, despite not having had genuine additional support adduced, because the believer is convinced that he has self-debiased. As Frantz (2006) reports, when experimental manipulation was used to induce a bias, "[g]iving

[6] John Randolph is a deejay and activist popularly known as *Jay Smooth*. His use of this metaphor was brought to my attention by Alexis Shotwell.

[7] http://www.bulletin.uwaterloo.ca/2009/aug/14fr.html.

participants instructions to be fair made them significantly less fair" (p. 165). The opportunity to introspectively ask "Am I being fair?" is an additional opportunity for bias to manifest, accompanied by a conviction that the answer is "Yes."

3 Successful and unsuccessful debiasing strategies

A longstanding literature on biases in cognitive and social psychology indicates a rather surprising array of intuitively sensible measures that don't work particularly well to mitigate various biases. Consider, for example, *hindsight bias*: the tendency to overestimate in retrospect one's confidence that the eventual actual outcome was antecedently the likeliest outcome. Subjects who rank the likelihood of outcomes before they take place (or before they are known) are disposed to subsequently misremember their rankings after they know the outcome, and to overrate the extent to which they predicted what actually took place. This is a bias of great interest and significance in its own right, because it bears so directly on the question of whether people are motivated to learn from their mistakes. A disposition to mentally sanitize the history of one's predictions is a disposition not to learn from mistaken predictions. Hence hindsight bias has attracted considerable attention from researchers in the heuristics and biases tradition of experimental psychology.

In his seminal work on debiasing, Baruch Fischhoff notes with wry amusement the extensive list of attempted remedies that showed little or no success in reducing hindsight bias among experimental subjects (1982, pp. 427–431). These include:

- Explicitly describing the bias to subjects and requesting that they avoid it
- Reducing the number of questions asked, to reduce cognitive load
- Inducing subjects to value the acuity of their memories of their predictions
- Asking subjects to estimate other subjects' foresight as well as their own
- Increasing the time between the report of the event and questioning the subjects
- Asking instead about the likelihood of recurrence, and not of the initial occurrence

With the first entry being the most obvious, there is something sensible and straightforward about each of these strategies; each of them proceeds from fairly transparent assumptions about the nature of learning, of cognition, or of error.

In this respect, the contrast between what doesn't work and what does work is significant and instructive. A debiasing measure that does show some efficacy in Fischhoff's analysis is for the subject to explicitly consider and entertain a range of alternative or counterfactual outcomes, and what would have had to happen in order for those outcomes to occur, before issuing a judgement on whether she had earlier rated the actual outcome as very likely (Fischhoff 1982, p. 430). So we do have at least one mitigation strategy with a significant prospect of success, taken as an experimental treatment. But how does it stack up as a debiasing measure outside the psychologist's lab? This is a crucial point from the perspective of making realistic plans to debias oneself, or to teach others how to debias themselves.

The approach in question is cited as a useful debiasing strategy by Wilson et al., who propose three kinds of practical self-administered strategies for debiasing in everyday life (Wilson et al. 2002, pp. 196–198).

1. Examining the reliability of one's judgements over multiple tests
2. Examining covariation of one's judgements with different biasing factors
3. Considering opposite or counterfactual scenarios

That these approaches are proposed for *everyday use* is the key point. In evaluating these strategies, we have not only to consider their effectiveness in experimental contexts, but their promise as strategies that can be self-administered by epistemic agents, in everyday situations, with the cognitive constraints and resource limitations that this implies.

Wilson et al. note that none of these three approaches is perfect, and that each is prone to failure. But there is a powerful distinction here that they do not emphasize. The first two strategies, but not the third, require that agents act like competent, dedicated and disinterested scientists to a considerable extent.[8] For many biases, including FP, to apply strategies (1) and (2) competently is to make and keep fairly extensive records of one's judgements, observationally coded in some perspicuous way, in order to measure differences and variations reliably. It is not a coincidence, I suggest, that the more persuasive examples Wilson et al. provide of people noticing their own biases by using the first two strategies involve the grading practices of their faculty colleagues (Wilson et al. 2002, pp. 196–197).[9] Such cases involve a structurally similar evaluative process, repeated many times over many years. They also involve a familiar metric for recording the judgements, with an independent professional requirement to keep records scrupulously.

But making, keeping, and interpreting records of one's judgements are not easy skills to learn, especially where the outcomes are hard to individuate or code. Expert scientists learn these skills over many years, apply them in formal contexts of mutual enforcement and correction, and still make plenty of mistakes. The idea that lay folk will manage to pull this off, working in isolation (under the influence of a biased disposition), is deeply implausible on its face. Most people are not psychology professors, and most instances of everyday bias do not involve assigning grades to student papers on a familiar conventional scale. FP in particular will be manifest in reactions and dispositions that the agent herself is not well-situated to notice, record and compare longitudinally. Indeed, it is far from obvious that professors or other cognoscenti, when they reason informally and outside the context of a research team, really are better at self-debiasing than most others (Berkowitz 1971; West et al. 2012). Hence I have a fairly deep skepticism about treating either the test-retest reliability of one's judgements or the covariation of responses with biasing factors as informal debiasing strategies.

What about the third strategy? Here we find an approach that looks less like an amateur research program in miniature, and more like a kind of self-initiated cognitive routine that could be undertaken preemptively, in context, when bias is a concern. Better still, this looks very much like the counterfactual scenario strategy, described by Fischhoff and reconfirmed by subsequent research, that significantly mitigates

[8] The first two strategies correspond loosely to J.S. Mill's Methods of Difference and Concomitant Variation.

[9] These examples also reveal a problem with treating self-debiasing as a process that people undertake once they have decided that they were *in fact* unduly biased (*pace* Wilson et al., pp. 187-189). Sometimes people run through a self-stimulated debiasing process merely because they think they *might* have been biased in a judgement, and they wish to check their reasoning.

hindsight bias. So the third strategy has some promise, both as an effective strategy when implemented in experimental contexts, and as a strategy that at least has some hope of actually being implementable in real-life situations.

4 Pessimism about debiasing false polarization

Of course, not every observation about what succeeds and fails in mitigating hindsight bias, or about debiasing in general, will apply to the FP case in particular. But we may nevertheless expect the most significant general lessons about debiasing to apply to FP as well for the most part. For example, when my interlocutor holds a position opposed to my own view on some emotionally fraught issue, it is intuitively appealing to suppose that my openness to fruitful dialogue will partially consist in my openness to the prospect that my own position is defeasible, or the possibility that my actual reasons fall short of being rationally compelling. This thought in turn might lead us to suppose that FP would be mitigated if I were to review my own arguments for cogency. By going through the exercise of articulating my own reasoning, on this hypothesis, I would promote conditions apt to help me appreciate the contrary view.

Analogous study of other biases suggests that these are mistaken assumptions and conclusions, and that such strategies—focusing on facts about the bias, reviewing my own reasons, and making mere mental efforts to avoid bias—will not prove effective. This suggestion appears correct. According to a study by Puccio and Ross (1998), partisans who articulated their *own* positions before testing were nevertheless found to be influenced by FP. By contrast, what seemed to work better was asking subjects to articulate the best arguments they could come up with for the *opposing* position.

> Partisans in the express-own-position condition in these studies showed the expected false polarization effect, markedly overestimating the gap between the positions of the two sides. By contrast, participants in the express-other-position condition (and, in one study, those in a third condition in which they expressed both positions) hardly overestimated this gap at all (Pronin et al. 2002a, p. 653).

In other words, it's not just that the naïve strategies that fail for a wide range of biases also fail for FP. What seems to mitigate FP also bears a close resemblance to what is seen to mitigate hindsight bias: namely, the strategy of reflecting not on actual outcomes or reasons, but on possible and counterfactual outcomes and reasons—ways things could have been, but aren't, and positions one could have held, but doesn't.

I have been at pains to note the respects in which debiasing is hard, and why this is so. Now we have at least a rough and ready candidate for an FP-debiasing strategy. Problem solved? Well, no. Debiasing is still extraordinarily hard, simply because the strategy of "considering the opposite" itself is very difficult to implement. But at least we can now discuss practical implementations of the strategy, and perhaps even read some characteristics of responsible social belief off those implementations.

What, then, is predictably difficult about the broadly counterfactual project of entertaining arguments you don't hold, or entertaining outcomes that didn't happen? More specifically, what problems arise when we aim to acquire or impart the skills to apply this strategy "in the wild", as self-policing epistemic agents? The answers here partly reflect the general problems canvassed earlier. I hypothesize three broad categories of

practical barriers to implementing a counterfactual scenario or "consider the opposite" approach to false polarization.

4.1 Agent-level launch problems for self-debiasing

In order for any bias-mitigating strategy to be effective, it has at least to be employed—and employed at the right times. Naturally, however, an agent's failing to recognize that she is biased in the first place means that the debiasing strategy will not be launched. The chief factor here is once again bias blindspot, which arises very specifically as a problem for FP.

We should not find the everyday significant of bias blindspot surprising, if we reflect on certain informal social phenomena. An example is polarization in argumentative exchanges between advocates of different religions, each of whom is nominally aware of the fact that religious views among the world's population are for the most part distributed *culturally* and *familially*, and not according to agents' broader evidence-evaluating abilities. That is, in many such discussions, disputants are surely aware that the factors determining most people's religious orientations are non-rational; yet they are convinced that their own views, at least, are supported by the most reasonable interpretation of the best available evidence. This looks very much like bias blindspot in action. Such informal observations are consistent with the empirical literature on false polarization, which reports partisans assuming that "their own views and assumptions were generally less shaped by political ideology than by objective or rational pragmatic concerns compared with the view of their adversaries or even their fellow partisans" (Robinson et al. 1995, p. 414).

Indeed, when we look at conjectured explanations for the bias blindspot, we can see why it would be particularly hard to overcome it in the case of FP. Pronin and Kugler (2007) suggest that bias blindspot is typically due to the *introspection illusion* that mental contents, accessible only from a first-person perspective, are better evidence of degree of bias than is behavior. They note, however, that mental contents are frequently unreliable, and are especially so in cases where one's judgements are already distorted by bias. Hence, because people have vastly greater access to their distorted introspective evidence in bias cases,

> people over-value thoughts, feelings, and other mental contents, relative to behavior, when assessing their own actions, motives, and preferences, but not when assessing others'. The term introspection illusion thus involves a self-other asymmetry in the relative valuation of introspective versus behavioral information (2007, p. 566).

I submit that this kind of informational asymmetery precisely characterizes the common individual dilemmas of Circle and Triangle giving rise to FP in Fig. 1. If the Pronin-Kugler explanation of bias blindspot in terms of introspection illusion is correct, we should expect FP to be one of the core cases of a bias vexed by blindspot issues. Irrespective of the correctness of that diagnosis, however, it is clear that bias blindspot comprises a significant launch problem for any FP-debiasing strategy. By undermining the perceived need to debias, it undermines the efficacy of strategies that can only be effective if they are used in the first place.

4.2 Agent-level implementation problems

We have already seen an agent-level implementation problem linked to the use of generally ineffective strategies: namely, that agents can easily form the perception of having already debiased, when in fact they have not. But even if we set aside introspective or otherwise naïve strategies, and concentrate on the express-other-position debiasing measure that seems effective in experimental contexts, there is prima facie reason to expect implementation problems arising from the agent's performance or attitudes. Why is this? Because evidence from other biases suggests that "consider the opposite" is what I will call a *brittle* strategy.

As Philip Tetlock notes with respect to hindsight bias, a dilemma lurks for those hoping to encourage the common use of a debiasing technique (Tetlock 2005, p. 199). On one hand, people unwilling to countenance the prospect that their judgements are mistaken will not avail themselves of the debiasing strategy; or, when they do, they will not be able to use it effectively. Even if a dogmatic overconfident agent can be induced to consider counterfactual scenarios in the hindsight bias case, her overconfidence is apt to manifest itself as an inability to call plausible alternative outcomes to mind. Thus the hindsight-biased agent who runs through an alternative-scenario debiasing exercise may simply confirm her original judgement that the actual outcome was the only live possibility. Believing herself correct—"I really am moderate, and my opponent really is an extremist! Why should *I* debias?"—she finds her bias influencing her very ability to formulate and entertain rationally defensible lines of support for her opponent's view.

By contrast, people willing to take their fallibility seriously may seem more epistemically virtuous. But those people are prone to overdoing the strategy, and thus may suffer an accuracy penalty for their openness to debiasing. The self-consciously fallible agent willing to consider alternative outcomes is apt to consider too many alternatives, and to find too many of them highly likely.[10] Either way, the strategy works best when it is guided by facilitators—as it typically is in experimental cases, to at least some degree. A reasonable role for an assistant is to help stimulate the generation of alternative scenarios when an agent is unimaginative, and impose probabilistic coherence on the generation of alternatives when the agent is imaginatively fecund or undisciplined. But with too few, too implausible, or too many counterfactuals in play, the strategy is apt to go awry in a variety of ways. Hence its brittleness; it is apt to break if bent in either direction.

4.3 Environmental implementation problems

Social, institutional, or other situational factors distinct from the specifically agent-level issues can also make it difficult to employ a "consider the opposite" approach

[10] Perhaps for similar reasons, O'Brien (2009) found it to be a more effective confirmation-debiasing technique to have criminal investigators consider just one alternative suspect than to have them consider three alternative suspects. The more alternatives are to be generated, with greater cognitive difficulty, the more apt one may be to regard the lower effort associated with the biased judgement as a kind of evidence in its favour (pp. 329–30).

to debiasing FP. Such factors entrench the misunderstandings and reasoning errors characteristic of false polarization. They arise from both groups that agree and groups that disagree with one's perspective.

What social reasons does Circle have for being slow to articulate or reflect upon the reasonable grounds that support Triangle's position, in an FP context? Circle might worry that any concession she makes will be taken as a sign of weakness—of mind, conviction, or character—by Triangle. Or she might fear having her own words used against her. Though Circle is a declared supporter of liberal abortion laws, she might nevertheless think that abortions are to be avoided if possible, other things being equal. But she will not savor the prospect of having Triangle, who espouses opposition to abortion, throw that concession back at her should she make it explicit. ("Even *you* admit that abortion is wrong!") Hence Circle might avoid such a potentially conciliating exercise. At the same time, the prospect of pressure from one's own camp can also inhibit the launch of bias-mitigating strategies. Fear of being perceived as disloyal, or as having abused the trust of her erstwhile co-believers by pretending agreement, may well motivate Circle to refrain from adopting the consider-the-opposite approach—except in a purely internal manner that may be limited by attentional resources and the effort or discomfort of harboring thoughts that cannot comfortably be expressed.

A different kind of environmental implementation problem is worth mentioning because it helps to explain why consider-the-opposite really does seem like the best bet for debiasing FP, even though the literature contains another FP-specific strategy. In fact the earliest diagnoses of FP arose from the literature on the social psychology of *negotiation*. In contexts of formal negotiation, a strategy quite different from consider-the-opposite was observed to be effective in mitigating FP: that of having negotiators on both sides make their commitments transparent in advance, *before* communications become strained (Keltner and Robinson 1993). Why not suggest this as an everyday FP-debiasing strategy in non-negotiation contexts as well?

The answer, in short, is that the structure of common discourse is quite different from that of controlled negotiation, and amounts to a powerful environmental constraint on the implementation of a debiasing strategy. For the most part, discussions, arguments, and the general uptake and interpretation of assertions in public discourse do not take the form of explicit negotiations. Partly for this reason, they often have no formally recognized moment of commencement of the sort that would enable one to employ the Keltner and Robinson strategy.

Even when a clear localized discussion does take place between two opposed interlocutors, so that we can identify some specific utterance as occupying discourse-initial position, this may not be the right level at which to individuate a perceived "opening move" in the debate. When emotionally charged topics like those characteristic of "culture wars" or "science wars" are in play, simply to discover that an interlocutor holds an opposing view may be to perceive the immediate discussion as part of an ongoing broader debate that has been underway for years. This is most likely what happened in the case described at the start of this paper.

More fundamentally, problematic cases of FP are apt to arise in aggregate from fragmentary or one-sided exchanges, so that no real discourse-initial utterance can easily be identified among the *actual* utterances. Given the sometimes uncomfortable nature of explicit, sustained discussions of a fraught topic, FP may well arise

between Artemis and Bill over the course of some months, during which time each of them has spoken to Catherine about the topic within earshot of the other, without either directly engaging the other on the topic. At least many FP cases, then, are sufficiently different from formal negotiation cases that no genuine opportunity will exist to present an itemized list of one's commitments before starting or pursuing the conversation. Since false polarization holds in situations that may already be characterized by strained communications, there is something to be said for a strategy that succeeds through one's generating and contemplating possible rationales for the opposing view, rather than depending on the contextual availability of an interlocutor's actual reasons. Negotiation-based strategies are not likely to work for many FP cases in non-negotiation contexts.

5 Optimism about debiasing false polarization

To summarize thus far: Debiasing is extraordinarily difficult for the most part. The approaches to it that one might intuitively expect to be effective have an alarming tendency either to be ineffective, or to worsen the bias. One broad strategy that is effective in experimentally mitigating many biases, including FP, seems quite difficult to implement in everyday contexts. People are likely either to think they needn't use it; or to use it poorly; or to recognize that they ought to use it, but refrain anyhow due to social pressures. The grounds for pessimism are extensive.

A natural thought, under these circumstances, is that we should just give up and settle for a cold beer and something good on television.

In fact there are at least modest grounds for optimism, and conclusions to be drawn from that optimism. One reason for optimism specific to FP is this: at least one major type of general debiasing error evaporates or is reduced in the case of FP. To see this, consider that typically three kinds of error may arise in a debiasing attempt: unnecessary correction, overcorrection, and insufficient correction. Wilson et al. conjecture that, of these three kinds, the least harmful is insufficient correction, since in every such case the corrected judgement is more accurate than the uncorrected one (Wilson et al. 2002, p. 191). Needless correction always leaves one worse off, while overcorrection could leave one better off or worse off.

This may be a broadly useful characterization, but it probably does not apply straightforwardly to FP. After all, the greatest possible overcorrection for FP would result in a discussant's believing that she and her interlocutor *do not really disagree at all*. This degree of overcorrection is inherently unlikely, but would in any case be self-correcting through subsequent discussion.[11] The procedural problem with FP is that it can hamper effective communication that would otherwise take place. But if

[11] There is reason to doubt that the extreme case of FP overcorrection will amount to the complementary error of the *False Consensus* (FC) effect—the overestimation of the extent to which one's views are shared by others, so named by Ross et al. (1977) in an early analysis of the phenomenon. FC is most strongly mediated by lack of evidence of what others actually think—for example, by the silence of interlocutors—while FP most easily arises from an (over-) interpretation of what interlocutors have said. The cognitive effort of interpreting a contrary utterance as an expression of complete agreement will plausibly constrain the prospect of overcorrection taking such an extreme form. FP bears important similarities to *pluralistic ignorance*, discussed in detail by Bicchieri and Fukui (1999), and by other work in this volume.

FP has been overcorrected, subsequent communication should be still less inhibited by perceived polarization, and this in turn should enable the points of disagreement between interlocutors to become clear. So it is not obvious that we should err on the side of undercorrection rather than overcorrection, when false polarization is the bias we are trying to mitigate.

The agent who overcorrects FP thereby enables sufficiently clear communication that a more accurate mutual understanding will eventually evolve from the situation. So even if willing debiasers have a tendency to overcorrect, the net effect should favour reliable belief-forming processes. This is particularly important, because it undermines one pessimistic thought: that the balance between undercorrection and overcorrection "in the wild" is so hard to manage that debiasing FP in everyday contexts will be practically impossible. The foregoing reflections give reason to think that there should be little worry about overdoing effective debiasing.

This sets the stage for a further optimistic thought. If consider-the-opposite is effective in experimental contexts, it may well be that we can implement some partial solutions to the problems canvassed in Sect. 4 by attempting to replicate or simulate some of these experimental conditions in everyday contexts. In the experimental contexts, for the most part, expert facilitators control substantial resources (infrastructure, apparatus, and personnel) in an effort to manipulate variables. They variously explain debiasing procedures, prompt subjects to undertake them, and may even guide them through the processes. In this way, subjects are walked through "artificially biased [i.e., consider-the-opposite] presentations" rather than being left to "unconstrained presentations of their actual views" (Pronin et al. 2002a, p. 653).

The sheer artifice of this effective FP-debiasing measure may seem to presage failure for the prospect of teaching—say, in a traditional university-level course on critical thinking, social epistemology, or cognitive science—a kind of debiasing recipe or algorithm for individuals to follow in their extra- and post-university lives. But a slightly different approach might well be effective. By focusing on inculcating the automatic or *habitual* launch of debiasing routines, for example, critical thinking education might partially address the bias blindspot worry. Education and training might also specifically emphasize the social permissibility of pointing out bias, as a means of mitigating barriers to the launch of debiasing strategies. It is unclear whether targeted education, over the long term, can undermine the intuitive plausibility of the misleading "mental effort" model of debiasing; but this too, if successful, would help address the false perception that one has already debiased. It is similarly unclear whether the effective use of consider-the-opposite, indulging its brittleness, is something that one can master through substantial practice and guided skill-development. The optimistic thought is that knowing what strategy could work, and why it is hard to implement, gives us a chance to target our educational and training decisions on facilitating the implementation of that strategy.

The most plausible approach, I suggest, is to make a range of efforts to create social infrastructure that supports debiasing. This might include institutionally and socially positioning facilitators who are trained to assist others in ways loosely analogous to the ways in which researchers assist subjects in effective trials of debiasing strategies. This sort of thing already happens to some degree, in some institutions and for some biases. The most obvious example is of judges in courts of law, who instruct jurors on

the unbiased interpretation of evidence, and who enforce the rules of evidence in the presentation of information during criminal trials. Similarly, some universities have expert consultants tasked to work with academic units in fields that traditionally have underrepresentation of women or of various races. These consultants advise decision-makers on ways to conceive of the student recruitment or faculty hiring processes, and on ways to structure those processes, in order to mitigate biases that may be at work in context (see, e.g., Bird et al. 2011, pp. 5–9).

Formal facilitation of this sort is expensive to implement and can slow down the reasoning process. It clearly could not be a *general* solution to everyday debiasing; we can't all walk around with a personal facilitator, just in case we find ourselves intractably disagreeing with someone! But that doesn't mean that societies or polities could not invest more heavily in the formal facilitation approach, applied in institutions where decisions matter most and where the potential for false polarization seems particularly high. The more familiar such facilitation processes were, arguably, the more habitual it might become even in the absence of facilitators themselves. A bootstrapping effect might be possible through informally socializing the role of facilitator, so that acquaintances or bystanders could be temporarily recruited for the role (or even volunteer, if they perceived polarization taking place). The potential for these outcomes to mitigate barriers to debiasing seems fairly clear. For example, when a partisan in an FP context is asked by a perceived-neutral third party to articulate some reasonable grounds for the opposing view, a range of launch problems are forestalled. The partisan has an immediate justification for doing so, besides that she is "giving up" or betraying the cause; so certain social barriers are plausibly diminished. I don't claim that multiplying the number and kind of social facilitators of debiasing would be easy or cheap. But I conjecture that it has some empirical possibility of working. And that is a relatively optimistic thought, all things considered.

To the extent that societies devote substantial resources to teaching reasoning skills, with the intention of imparting enhanced critical thinking and belief-forming abilities or dispositions, the evidence is that these are highly valued social and education outcomes. There is a fairly clear specific value, moreover, to avoiding false polarization that shuts down the fruitful exchange of views. The foregoing data and reflections suggest that debiasing in general, including debiasing socially distorting effects that lead to belief polarization, is a much harder part of cultivating reliable belief-forming processes than one might have suspected. Indeed, it is so difficult that perhaps no standard classroom pedagogy will be really effective in enabling it.

But if anything will work, the evidence suggests the working hypothesis that it will be some combination of extensive targeted training, the inculcation of certain debiasing habits, a broad spectrum of efforts to impart social attitudes that tolerate and respect conciliating and changing one's mind, and the creation of dedicated reasoning infrastructure, including expert assistants, to guide debiasing efforts in contexts where they are likely to arise. In one practical sense, this provides some useful cues for how to structure, or restructure, the industry of teaching strong reasoning skills. But in another sense, it reveals the extent to which the projects of acquiring and imparting the skills of effective reasoning and communication are profoundly socialized projects. There is no realistic hope of cultivating the relevant kinds of reliable belief-forming processes without a vast network of background educational support structures, and

probably not without a great deal of ongoing social and institutional reasoning support as well. Reasoning well, overcoming biases, and remedying belief polarization are socially epistemological concerns not merely in the sense that analysis of their concepts requires focusing on the role of interacting agents. They are moreover generally achievable only through socially cooperative commitment, investment, and implementation. There is little plausibility to an atomistic conception of education, or of agents, on which the goal of critical thinking pedagogy is to create self-sufficient epistemic agents who can self-correct for biases of this sort—or of practically any sort.[12]

References

Berkowitz, L. (1971). Reporting an experiment: A case study in leveling, sharpening and assimilation. *Journal of Experimental Social Psychology*, *72*, 237–243.

Bicchieri, C., & Fukui, Y. (1999). The great illusion: Ignorance, informational cascades, and the persistence of unpopular norms. In M. Galavotti & A. Pagnini (Eds.), *Experience, reality, and scientific explanation* (pp. 89–121). Dordrecht: Kluwer.

Bird, S. R., Fehr, C., Larson, L. M., Sween, M. (2011). *ISU ADVANCE Collaborative transformation project: Final focal department synthesis report*. Iowa State University ADVANCE Program. Report available online at: http://www.advance.iastate.edu/resources/resources.shtml.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge: Cambridge University Press.

Frantz, C., & Janoff-Bulman, R. (2000). Considering both sides: The limits of perspective-taking. *Basic and Applied Social Psychology*, *22*, 31–42.

Frantz, C. (2006). I AM being fair: The bias blind spot as a stumbling block to seeing both sides. *Basic and Applied Social Psychology*, *28*(2), 157–167.

Goldman, A. (1986). *Epistemology and cognition*. Cambridge, MA: Harvard University Press.

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, *69*, 1069–1086.

Keltner, D., & Robinson, R. (1993). Imagined ideological differences in conflict escalation and resolution. *The International Journal of Conflict Management*, *4*, 249–262.

Keltner, D., & Robinson, R. (1996). Extremism, power, and the imagined basis of social conflict. *Current Directions in Psychological Science*, *5*, 101–105.

Kennedy, K. A., & Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin*, *34*, 833–848.

Lilienfeld, S., Ammirati, R., & Landfield, K. (2009). Giving debiasing away. *Perspectives on Psychological Science*, *44*, 390–8.

Monin, B., & Norton, M. (2003). Perceptions of a fluid consensus: Uniqueness bias, false consensus, false polarization, and pluralistic ignorance in a water conservation crisis. *Personal and Social Psychology Bulletin*, *295*, 559–67.

O'Brien, B. (2009). Prime suspect: An examination of factors that aggravate and counteract confirmation bias in criminal investigations. *Psychology, Public Policy, and Law*, *154*, 315–334.

Pronin, E., Puccio, C., & Ross, L. (2002a). Understanding misunderstanding: Social psychological perspectives. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristic and biases: The psychology of intuitive judgement* (pp. 636–665). Cambridge: Cambridge University Press.

Pronin, E., Lin, D., & Ross, L. (2002b). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*, 369–381.

Pronin, E., & Kugler, M. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *434*, 565–578.

[12] For helpful comments on earlier drafts I am grateful to Guillaume Beaulac, Gerald Callaghan, Carla Fehr, Carlo Proietti, Frank Zenker, fellow participants in the April 2011 Copenhagen/Lund Workshop in Social Epistemology, and two anonymous referees for this issue. This work was supported in part by the Faculty of Arts, University of Waterloo, and by Social Sciences and Research Council of Canada Grant 410-2011-1737.

Puccio, C., & Ross, L. (1998). Real versus perceived ideological differences: Can we close the gap?. Unpublished ms, Stanford University

Randolph, J. (2012). How I learned to stop worrying and love discussing race. http://tedxtalks.ted.com/video/TEDxHampshireCollege-Jay-Smooth. Accessed April 20, 2012.

Robinson, R., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naive realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, *68*, 404–417.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301.

Sanna, L., Stocker, S., & Schwarz, N. (2002). When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 497–502.

Strayer, D., Drews, F., & Johnston, W. (2012). The eye of the beholder: Cellular communication causes in attention blindness behind the wheel. In A. G. Gale (Ed.), *Vision in vehicles X* (pp. 142–148). Applied Vision Research Centre, Loughborough University.

Tetlock, P. (2005). *Expert political judgment*. Princeton, NJ: Princeton University Press.

Thompson, L. (1995). "They saw a negotiation": Partisanship and involvement. *Journal of Personality and Social Psychology*, *68*, 839–853.

West, R., Meserve, R., & Stanovich, K. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*. Advance online publication. doi:10.1037/a0028857

Willingham, D. (2007). Critical thinking: Why is it so hard to teach? *American Educator*, *31*(2), 8–19.

Wilson, T., Centerbar, D., & Brekke, N. (2002). Mental contamination and the debiasing problem. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristic and biases: The psychology of intuitive judgement* (pp. 185–200). Cambridge: Cambridge University Press.