On overrating oneself... and knowing it*

Adam Elga

December 11, 2003
Revised version to appear in *Philosophical Studies*

Abstract

When it comes to evaluating our own abilities and prospects, most (non-depressed) people are subject to a distorting bias. We think that we are better—friendlier, more well-liked, better leaders, and better drivers—than we really are. Once we learn about this bias, we should ratchet down our self-evaluations to correct for it. But we don't. That leaves us with an uncomfortable tension in our beliefs: we knowingly allow our beliefs to differ from the ones that we think are supported by our evidence. We can mitigate the tension by waffling between two belief states: a reflective state that has been recalibrated to take into account our tendency to overrate ourselves, and a non-reflective state that has not.

My friend Daria believed in astrology. For example, she thought that because of her astrological sign she was going to be particularly lucky over the next few weeks. That was bad enough. But when I tried to persuade her that astrology is unfounded, I discovered something even worse.

I gave Daria evidence against astrology—studies showing that the position of the distant stars at the time of one's birth has no bearing on one's personality or prospects. Daria agreed that the studies were significant evidence against the truth of astrology, and that she had no countervailing evidence of comparable strength. But that was not the end of the matter. "I still believe in astrology just as much as I did before seeing the studies," she said. "Believing in astrology makes me happy."

*<adame@princeton.edu>. Thanks to Emily Pronin, Michael Fara, Ishani Maitra, Dmitri Tymoczko, Lucy Harman, Jordan Lite, the 2002 Old Dominion Fellows, the attendees of the 2003 Bellingham Summer Philosophy Conference, and an anonymous referee for helpful feedback, and to the Princeton University Humanities Council for research support.

I was floored. Daria's original belief in astrology was less than perfectly reasonable. But this—believing in astrology even though *by her own lights* the evidence went against it—was an insult to rationality. And it was no excuse that her belief in astrology made her happy.

Daria's original belief in astrology was, I think, a product of wishful thinking. Because she wanted to believe in astrology, she was unconsciously biased in favor of it. For example, she would attend more carefully to instances in which astrological predictions came out right, than ones in which they came out wrong. It was unreasonable for Daria to attend to her evidence in this biased way. But at least she ended up with what, by her lights, were good reason for her beliefs.[1]

Not so for her belief in astrology *after* learning about the anti-astrology studies. For then her beliefs went against what she thought the evidence supported. She learned about the studies, and thought that they were evidence against astrology. But—with that fact firmly in mind—she believed in astrology just as much as before.

Daria knowingly violated the following norm of rationality:

(E) One ought not have beliefs that go against what one reasonably thinks one's evidence supports.

(If you doubt that this is a norm, you can make its plausibility vivid by imagining an argument with an opponent who violates it. You marshal your best evidence for your view. Your opponent agrees that you've presented strong evidence for your view, and has no counter-evidence. But no matter how much evidence you present, or how strong it is, he gains no confidence in your view. I invite you to agree that in this infuriating scenario, your opponent is being unreasonable.[2])

The case of Daria raises several theoretical problems. First, it seems as though Daria *deceives herself* about astrology. One might wonder what exactly it takes to deceive oneself in this way. Second, one might wonder how such self deception is possible in the first place. I will not pursue these questions. I will simply assume that people can and do deceive themselves in the way that Daria does, even though it is no easy task to say how they manage to do so (Bach 1981, Johnston 1988, Scott-Kakures 1996).

My target is not a theoretical problem, but a practical one. The problem is: how can one square one's commitment to (E) with what appear to be one's

[1] On the distinction between wishful thinking and more extreme forms of self-deception, see Bach (1981), Scott-Kakures (2000).

[2] Note that your opponent is still being unreasonable even if he has strong practical motivations to hold on to his view.

own knowing and persistent violations of it?

I faced the problem when I became convinced that I violated (E) even more than Daria did. If you are convinced by the evidence below, then you may begin to think that you violate (E), too.

1

I began thinking that I knowingly and persistently violated (E) after I read some social psychology studies about self-evaluations.

It turns out that people have inflated views of their own abilities and prospects. People (nondepressed people, at least) rate themselves as better—friendlier, more likely to have gifted children, more in control of their lives, more likely to quickly recover from illness, less likely to get ill in the first place, better leaders, and better drivers—than they really are. And that's just the beginning. There is a great deal of work documenting the persistent and widespread positive illusions (about themselves) to which people are subject.[3,4]

In contrast, depressed people have been found to have more accurate self-evaluations.[5] That accuracy probably doesn't help them. There is evidence associating the above sorts of positive illusions with increased happiness, "ability to care for others", "motivation, persistence", and "the capacity for creative, productive work" (Taylor and Brown 1988). Furthermore, there is evidence that at least some of the association is causal: that positive illusions help people get by.

None of this is surprising. It is reasonable to think that normal (i.e., unrealistically high) self-evaluations promote the sort of self-esteem and self-confidence that help people start projects and persist through difficulties. And it is reasonable to think that a positive self-image makes people happier.

---

[3]Here and below I rely on Taylor and Brown (1988), Taylor and Brown (1994), Brown (1986), and Lehman and Taylor (1987).

[4]The above description suggests that people overrate themselves in *every* respect, which isn't quite true. But people do overrate themselves in a great many respects: "on virtually every conceivable positively valued trait, the majority of people think that they are better than others" (Brown and Dutton (1995), as cited in Scott-Kakures (2000)). Furthermore, the discussion to follow would still apply if we restricted attention to just those respects in which people have been found to overrate themselves.

[5]Note that the evidence for this is weaker: see Taylor and Brown (1994) and Reed et al. (1994).

2

How should one respond to the above evidence? In particular, how should one change one's self-evaluations? When I first faced that question, I was a hard-liner about Daria, so the answer was clear: the only acceptable response was immediate and total recalibration. Just as Daria should have decreased her confidence in astrology once she heard about the anti-astrology studies, so I should have consistently judged myself to be less friendly, in control, and all the rest, once I read the positive illusion literature.

It didn't happen.

I was convinced that most people overrate themselves, and had no reason to think I was an exception. I mouthed the words "I'm not as good as I thought I was." But they didn't sink in. As soon as it was time to make dinner, write a paper, or see a friend—indeed, as soon as it was time to do anything but sit in my office brooding about the positive illusion literature—the impact of that literature on my self-evaluations completely evaporated.

Try it yourself. If you were at all convinced by the above summary of the positive illusion literature, see if it lowered your estimate of how good a writer you are. How good a lover. How likely you are to get tenure, or some distinguished chair.

It is tough to make a sustained change in one's self-evaluations. Just learning that people overrate themselves does not automatically effect such a change. The same is not true for all judgments. For example, if I were to become convinced that people tend to overrate the time they spend in elevators, I'd have no trouble reducing my estimate of how long I had recently spent in elevators. Indeed, I don't think I'd have any choice in the matter. Just finding out about the elevator bias would change my beliefs, period. In contrast, it is easy to continue to overrate oneself even after learning that people are generally prone to doing so.

(Unsurprisingly, people's positive illusions persist even when they are told about the prevalence of such illusions. For example, Pronin et al. (2002) and Friedrich (1996) have run experiments in which subjects are explicitly informed about the tendency to overrate oneself. The result: subjects acknowledge that *others* have inflated self-evaluations, but insist that their *own* self-evaluations have been realistic or "overly modest" (Pronin et al. 2002, p. 375).)

In sum: I counted the positive illusion literature as evidence that, like most people, I tend to overrate myself. But this evidence had no effect on my everyday self-evaluations. As a result, it seemed that I was a knowing and persistent violator of (E) (the norm that one's beliefs ought not go against what one reasonably thinks one's evidence supports).

Given my strong endorsement of (E), the thought that I was such a violator presented me with a practical problem—the problem of squaring my practices with my epistemic conscience. Furthermore, the problem didn't arise from quirky facts about me. A similar problem faces every friend of (E) who becomes convinced that she is subject to positive illusions about herself, but who does not as a result lower her self-evaluations.

3

One way around the problem would be to find an escape route—a reason for thinking that the positive illusion literature is no evidence that one overrates oneself. With such an escape route one could remove all pressure to downgrade one's self-evaluations. But the three most plausible routes do not stand up to scrutiny.

Escape route 1: "The positive illusion evidence is merely statistical. True, people overrate themselves in general. But that gives me no reason to downgrade my self-evaluation in some particular respect on some particular occasion."[6]

Reply: Statistical evidence of this sort *should* influence one's self-evaluations. For comparison, consider the case of a pilot who gets statistical evidence that the altimeters in planes tend to indicate overly high altitudes. If he gets the news while in the air (and if he has been relying on his altimeter to judge his altitude), the news should make him lower his estimate of his altitude.

Escape route 2: "People tend to count themselves as better than average not because they overrate themselves, but rather because they rate themselves using criteria favorable to their own strengths. For example, cautious drivers may count themselves as excellent drivers because of their safe practices, while fast drivers may count themselves as excellent drivers because of their decisiveness and efficiency."[7]

Reply: This effect is certainly responsible for *some* of the observed results (Dunning et al. 1989). But it does not explain why people overrate themselves even with respect to fixed criteria. For example, it does not explain why people overestimate the speed at which they will recover from illness.

Escape route 3: "Granted, the subjects in social psychology studies are subject to positive illusions. But I am a member of a group which is not well represented by the subjects of those studies. So the studies don't apply to me."

[6] I am grateful to Juan Comesaña for bringing this potential escape route to my attention.

[7] I am indebted to Emily Pronin for drawing this reply to my attention.

Reply: It is true that many such studies have used as subjects undergraduates at psychology research institutions. But the illusions have also been found in quite diverse subject pools. For example, managers have been found to overrate their managerial abilities.[8] Another example: "94% of college professors say they do above-average work."[9] Though I know of no studies targeting philosophers in particular, there is every reason to think that philosophers are subject to the illusions. Indeed, there is reason to think that philosophers are particularly vulnerable. For one of the most persistent illusions is that one's own views—more than the views of one's peers—have been formed by objectively evaluating the weight of the evidence (Pronin et al. 2003).

4

The escape routes above don't work. So I am left with the problem of reconciling my own harsh epistemic criticisms of Daria with the fact that I seem to be subject to those very same criticisms. What follows is my proposed reconciliation.

I am of two minds about my own abilities and prospects. In my moments of coolest rational reflection, when I am staring the positive illusion evidence right in the face, I *do* lower my self-evaluations in the light of that evidence.

When I enter the fray, however, these considerations lose their influence. It is not that I forget about them completely. It is rather that in ordinary life—in deciding what to eat for breakfast, say, or whether to shoot or pass the basketball—certain considerations play a serious role, and others get shoved on the back burner. The positive illusion evidence (and the reasoning that leads from that evidence to recalibrating my self-evaluations) gets shoved on the back burner.[10] That all happens automatically, and probably is accomplished by the same mechanisms that created the positive illusions in the first place.[11]

So when it comes to self-evaluation, I waffle between two belief states. My *reflective* belief state takes into account the positive illusion literature, and

[8] Larwood and Whittaker (1977), as cited in Dunning et al. (1989).

[9] Cross (1977), as cited in Dunning et al. (1989, p. 1082).

[10] I owe this suggestion to David Lewis. Compare it to Hume's famous observation that his "philosophical melancholy" could not be neutralized by reasoned argument, but only by backgammon and merriment (Hume 1738).

[11] These mechanisms—instances of what Johnston (1988) calls "mental tropisms"—include: remembering past successes more than failures, seeing one's own performance at tasks as being better than it is, and counting the respects in which one is strong as more important than the respects in which one is weak.

my *non-reflective* one does not. That setup isn't perfectly reasonable. But I'd like to point out a respect in which Daria is even less reasonable.

Talk to Daria. Have her explain the anti-astrology studies. Get her to admit that even by her own lights, these studies count as strong evidence against astrology. Suppose that she does so. Suppose that while fully reflecting on all of the above, her belief in astrology remains just as strong as before she learned about the studies. Then Daria is being unreasonable. Her beliefs at that time don't fit together properly.

It is not that Daria's beliefs are inconsistent. She does not both believe that astrology is correct and incorrect. No—when it comes to the question of whether astrology is correct, her opinion is clear: she is confident that astrology is correct. It's just that she *also* believes that she has strong *evidence* that astrology is incorrect, and not much countervailing evidence.

Daria has beliefs that—by their own lights—go against the evidence. And this combination of beliefs persists even when she is aware of the tension. Contrast Daria with someone who waffles—someone who has a reflective belief state that takes into account the positive illusion literature, and a non-reflective one that does not. The waffler does not suffer from the above failing of rationality. In his reflective state, he takes into account that his perceptions of his abilities tend to be positively biased, and adjusts them accordingly. No failure there.

In his non-reflective state, there is a latent tension in his beliefs: he knows about positive illusions, but is acting in a way that doesn't take into account of that knowledge. But (unlike Daria) he is disposed to properly resolve that tension when he attends to it. If in the middle of giving a lecture (while he is thinking "This is going great!"), he were forced to take a break and reevaluate how well the lecture was going (in the light of the positive illusion literature), he *would* recalibrate. He would admit that the lecture probably wasn't going as well as he'd thought it was.

In sum: both Daria and I violate (E)—the norm that one's beliefs ought not go against what one thinks one's evidence supports. But Daria violates it always, even when that very violation is brought to her attention. In contrast, I violate the norm only when in my non-reflective state. When the violation is brought to my attention, I am disposed to recalibrate in order to eliminate it. So: it is true that I am a persistent violator of the norm, and know this about myself. But at no time do I *both* recognize that I am violating the norm, *and* persist in violating it. I suggest that the same may be true for others who find the positive illusion literature convincing. Recognizing this should help such people square their endorsement of the norm with their inability to fully abide by it.

References

Kent Bach. An analysis of self-deception. *Philosophy and Phenomenological Research*, 41(3):351–370, 1981.

J. Brown and K. Dutton. Truth and consequences: the costs and benefits of accurate self-knowldege. *Journal of Personality and Social Psychology*, 21: 1288–1296, 1995.

Jonathon D. Brown. Evaluations of self and others: self-enhancement biases in social judgments. *Social Cognition*, 4(4):353–376, 1986.

P. Cross. Not can but will college teaching be improved. *New Directions for Higher Education*, pages 1–15, 1977.

David Dunning, Judith A. Meyerowitz, and Amy D. Holzberg. Ambiguity and self-evaluation: the role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, 57(6): 1082–1090, 1989.

James Friedrich. On seeing oneself as less self-serving than others: the ultimate self-serving bias? *Teaching of Psychology*, 23(2), 1996.

David Hume. *A treatise of human nature*. Clarendon Press, Oxford, 1738. Reprint 1966.

Mark Johnston. Self-deception and the nature of mind. In Brian P. McLaughlin and Amelie Oksenberg Rorty, editors, *Perspectives on self-deception*, pages 63–91. University of California Press, Berkeley, 1988.

L. Larwood and W. Whittaker. Managerial myopia: Self-serving biases in organizational planning. *Journal of Applied Psychology*, 62:194–198, 1977.

Darrin R. Lehman and Shelley E. Taylor. Date with an earthquake: Coping with a probable, unpredictable disaster. *Personality and Social Psychology Bulletin*, 13(4):546–555, 1987.

Emily Pronin, Thomas Gilovich, and Lee Ross. Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. *Psychological Review*, 2003. Forthcoming.

Emily Pronin, Daniel Y. Lin, and Lee Ross. The bias blind spot: perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, 28(3): 369–381, 2002.

Geoffrey M. Reed, Margaret E. Kemeny, Shelley E. Taylor, Hui-Ying J. Wang, and Barbara R. Visscher. Realistic acceptance as a predictor of decreased survival time in gay men with aids. *Health Psychology*, 13(4):299–307, 1994.

Dion Scott-Kakures. Self-deception and internal irrationality. *Philosophy and Phenomenological Research*, 56(1), 1996.

Dion Scott-Kakures. Motivated believing: wishful and unwelcome. *Nous*, 34 (3):348–375, 2000.

Shelley Taylor and Jonathon Brown. Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103(2):193–210, March 1988.

Shelley Taylor and Jonathon Brown. Positive illusions and well-being revisited: separating fact from fiction. *Psychological Bulletin*, 116(1):21–27, July 1994.