# RETHINKING THE DEBRIEFING PARADIGM: THE RATIONALITY OF BELIEF PERSEVERANCE

David M. GODDEN

ABSTRACT: By examining particular cases of belief perseverance following the undermining of their original evidentiary grounds, this paper considers two theories of rational belief revision: foundation and coherence. Gilbert Harman has argued for coherence over foundationalism on the grounds that the foundations theory absurdly deems most of our beliefs to be not rationally held. A consequence of the unacceptability of foundationalism is that belief perseverance is rational. This paper defends the intuitive judgement that belief perseverance is irrational by offering a competing explanation of what goes on in cases like the debriefing paradigm which does not rely upon foundationalist principles but instead shows that such cases are properly viewed as instances of positive undermining of the sort described by the coherence theory.

KEYWORDS: belief perseverance, belief revision, debriefing paradigm, bounded rationality, coherence theory, foundationalism, principle of positive undermining, rationality

## 1. Introduction

The phenomenon of belief perseverance, which occurs when beliefs survive "the total destruction of their original evidential basis,"[1] presents at least two problems for the theory of reasoning and rationality. First is the descriptive and psychological problem of describing the nature and extent of the phenomenon, and of explaining how and why it occurs. Second is the normative and epistemological problem of whether, and to what extent, belief perseverance is rational. This paper concerns the second of these problems.

Typically belief perseverance is viewed as failure of rationality on the part of the reasoner.[2] For example, Ross, Lepper and Hubbard described the effect of their

[1] Lee Ross, Craig A. Anderson, "Shortcomings in the Attribution Process: On the Origins and Maintenance of Erroneous Social Assessments," in *Judgment under Uncertainty: Heuristics and Biases,* eds. Daniel Kahnemann, Paul Slovic, and Amos Tversky (Cambridge: Cambridge University Press, 1982), 149.

[2] Craig A. Anderson, "Belief Perseverance," in *Encyclopedia of Social Psychology*, eds. Roy Baumeister and Kathleen D. Vohs (Thousand Oaks: Sage, 2007), 109-110; Richard E Nisbett, Lee

debriefing paradigm (in which evidence demonstrating to the satisfaction of a reasoner that the original evidential basis for one of her beliefs is completely unfounded) as having "far less impact [on the reasoner's attitude to the resultant belief] than would be demanded by any logical or rational impression-formation model."[3] This intuitive view has prompted theorists of reasoning to classify belief perseverance as a cognitive bias along with phenomena like the confirmation bias,[4] the conjunction fallacy,[5] and the belief bias.[6] Indeed, it is primarily because belief perseverance is deemed to be irrational that it presents theoretical problems for accounts of human rationality and moral problems for experimenters using deception and debriefing paradigms in psychological research.

Against this view, Gilbert Harman[7] has claimed that belief perseverance is not irrational. Harman argues that the rational condemnation of belief perseverance relies on a foundationalist epistemology and theory of rational belief change. Foundationalist theories involve a principle of negative undermining which requires that subjects track all of the reasons they have for their beliefs, and make rationally appropriate adjustments to (their confidence levels in) their beliefs whenever the basing relations among them changes. Yet, Harman argues, in view of the cognitive limitations of normal human reasoners, the foundationalist theory of rational belief change is not consistent with the principles of bounded rationality. Indeed, Harman argues, on a foundationalist account most of our beliefs are not rationally held. Because of this Harman claims that the foundationalist theory of

---

Ross, *Human Inference: Strategies and Shortcomings of Social Judgment* (Englewood Cliffs: Prentice-Hall, 1980); Craig A. Anderson, Mark R. Lepper, and Lee Ross, "Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information," *Journal of Personality and Social Psychology* 39 (1980): 1037-1049.

[3] Lee Ross, Mark R. Lepper, and Michael Hubbard, "Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm," *Journal of Personality and Social Psychology* 32 (1975): 880.

[4] Peter C. Wason, "Reasoning," in *New Horizons in Psychology*, ed. Brian M. Foss (Harmondsworth: Penguin, 1966), 135-151; Peter C. Wason, "Reasoning About a Rule," *Quarterly Journal of Experimental Psychology* 20 (1968): 273-281.

[5] Amos Tversky and Daniel Kahneman, "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychological Review* 90 (1983): 293-315.

[6] Jonathan St. B. T. Evans, Julie L. Barston, and Paul Pollard, "On the Conflict Between Logic and Belief in Syllogistic Reasoning," *Memory and Cognition* 11 (1983): 295-306.

[7] Gilbert Harman, *Change in View: Principles of Reasoning* (Cambridge: The MIT Press, 1986); Gilbert Harman, "Internal Critique: A Logic is not a Theory of Reasoning and a Theory of Reasoning is not a Logic," in *Handbook of the Logic of Argument and Inference: The Turn Towards the Practical*, eds. Dov M. Gabbay, Ralph H. Johnson, Hans J. Olbach, and John Woods (New York: Elsevier, 2002), 171-186.

rational belief change cannot be correct. As a corollary, belief perseverance is rational.

In place of the foundations theory Harman advocates for the coherence theory of rational belief change. The coherence theory offers an account of what goes on in cases like the debriefing paradigm which renders the belief perseverance behavior rational. This explanatory 'success' is then counted as evidence for the normative correctness of the coherence theory.

In this paper, I defend what I take to be our intuitive judgement that belief perseverance is indeed irrational. I do this by offering a competing explanation to Harman's own which classifies belief perseverance as it occurs in the debriefing paradigm as irrational without relying upon foundationalist principles. The account thereby avoids the controversial and putatively unacceptable consequence that the majority of our beliefs are not rationally held.

## 2. Rationality & Bounded Rationality

Reasoning (or inference) is a psychological process of reasoned change in view,[8] or belief revision, which involves "trying to improve one's overall view by adding some things and subtracting others."[9] The goal one aims at when improving one's overall view is rationality, and it is against standards of rationality that one's overall view, and the revisions made to it, are measured. It is here that epistemology and logic contribute to the theory of rational belief revision.

Historically, the normative study of rationality began with the specification of a formal system thought to embody a set of rational ideals. Judgements of rationality in individual cases were then made according to whether and how behavior satisfied the requirements of the formal system.[10] Yet, as Chater and Oaskford have observed,[11] there is something paradoxical about research in the normative qualities of human reasoning. Not only does any attempt to assess the reliability of human reasoning involve, indeed rely upon, the very processes whose reliability we are attempting to assess. But any standards we seek to provide as norms of good reasoning will themselves be products of human reasoning processes. Further, what is to be said of otherwise rational agents who regularly seem in default of the prescribed standard?

---

[8] Harman, "Internal Critique," 171.

[9] Gilbert Harman, "Logic, Reasoning, and Logical Form," in *Language, Mind and Brain*, eds. Thomas W. Simon and Robert J. Scholes (Hillsdale: Lawrence Erlbaum Associates, 1982), 13.

[10] Ken Manktelow, *Reasoning and Thinking* (Hove: Psychology Press, 1999), 5.

[11] Nick Chater and Mike Oaksford, "Human Rationality and the Psychology of Reasoning: Where Do We Go From Here?" *British Journal of Psychology* 92 (2001): 193-216.

Cohen argued that empirical studies cannot contribute to a demonstration of systematic irrationality in human agents.[12] Rather, Cohen argued, in order to conduct any empirical investigation into the success of human reasoning "humans have to be attributed a competence for reasoning validly, and this provides the backcloth against which we can study defects in their actual performance."[13] In the same vein, thinkers such as Dennett[14] have advanced a position which Stich called the *argument from the inevitable rationality of believers*.[15] On Stich's reconstruction, Dennett does not hold that people must be rational, but that "people must be rational *if they can usefully be viewed as having any beliefs at all*," or as Stich puts it "intentional descriptions and rationality come in the same package."[16] Against these views Stich has argued that *a priori* arguments seeking to show that human irrationality cannot be empirically demonstrated are miscast.[17]

My own view is roughly that of Peirce, as set out in "The fixation of belief," where he maintained that we are "in the main logical animals, but we are not perfectly so."[18] Perhaps another way to state this type of position, more in line with the notion of bounded rationality and evolutionary epistemology, can be found in Nisbett and Ross's thesis that "people's inferential strategies are well adapted to deal with a wide range of problems, but that these same strategies become a liability when applied beyond that range."[19]

### 2.1. Bounded Rationality

Yet, there is something to Cohen's view that our rational norms must be competence norms. If a normative theory of reasoning is to be prescriptive over our actual inferential practices, then it seems as though we must be able to follow the prescriptions made by the theory. So, a general problem for theorists seeking to provide a normative theory of reasoning stems from our nature as rational agents with a finite cognitive endowment.

A cardinal example of this is the idea of deductive closure: that the beliefs of a perfectly rational agent should be closed under the laws of deduction. A

---

[12] Jonathan L. Cohen, "Can Human Irrationality Be Experimentally Demonstrated?" *Behavioral and Brain Sciences* 4 (1981): 317-370.

[13] Cohen, "Can Human Irrationality," 317.

[14] Daniel Dennett, *Brainstorms* (Montgomery: Bradford Books, 1978).

[15] Stephen Stich, "Could Man Be An Irrational Animal?" *Synthese* 64 (1985): 120.

[16] Stich, "Could Man Be," *Synthese* 64 (1985): 121.

[17] Stich, "Could Man Be," Stephen Stich, *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation* (Cambridge: The MIT Press, 1990).

[18] Charles Sanders Peirce, "The Fixation of Belief," *Popular Science Monthly* 12 (1977): 1-15.

[19] Nisbett and Ross, *Human Inference*, xii.

consequence of this principle is that agents should believe all of the logical consequences of their present beliefs. Yet, even a single belief, *p*, generates an infinitude of logical consequences through the law of disjunction introduction (*p* |- *p* V *q*). Thus, any single belief, *p*, would generate the following infinite series of logical consequences: *p* V *q*, (*p* V *q*) V *r*, [(*p* V *q*) V *r*] V *s*, …, etc. As Harman rightly points out, not only is it impossible for a finite cognitive agent to follow through on all of the consequences of her present beliefs, it is often neither practical nor prudent to even begin to do so.[20] With this in mind, Harman introduces the principle of *clutter avoidance* – that "one should not clutter one's mind with trivialities" – as an example of the type of principle that belongs in a properly articulated theory of rational belief change. [21]

It is not just that the principles of ideal rationality do not apply to agents whose powers of reasoning are finite; they also do not apply to rational agents whose judgement is fallible. Take, for example, the principle of consistency: that all of our beliefs should be consistent with one another. Using an example like the preface paradox,[22] Harman observes not only is it not possible to attain perfect consistency among all of our beliefs, for fallible judges it is not rational to have perfectly consistent beliefs. He writes:

> a rational fallible person ought to believe that at least one of his or her beliefs is false. But then not all of his or her beliefs can be true, since, if all of the other beliefs are true, this last one will be false. So in this sense a rational person's beliefs are inconsistent. It can be proved they cannot be all true together.[23]

Considering this type of case, Pinto argues (i) that so long as one does not use contradictory premises to make inferences, that the rational fault of having inconsistent beliefs can be no more serious than that of having a false belief,[24] and (ii) that unless one has a particular reason to suspect one of the beliefs involved in the inconsistency is dubious, it is not rational to give any of them up upon discovering an inconsistency.[25] Thus, it should be recognized at the outset that the ideals of perfect rationality simply do not apply in any direct fashion to cognitively

[20] Harman, *Change in View*, 12.
[21] Harman, *Change in View*, 12.
[22] See David C. Makinson, "The Paradox of the Preface," *Analysis* 25 (1965): 205-207.
[23] Gilbert Harman, "Logic and Reasoning," *Synthese* 60 (1984): 107.
[24] Robert C. Pinto, "Inconsistency, Rationality and Relativism," in *Argument, Inference and Dialectic: Collected Papers on Informal Logic*, eds. Robert C. Pinto and Hans V. Hansen (Dordrecht: Kluwer, 2001), 46-51.
[25] Pinto, "Inconsistency, Rationality," 51-53.

finite rational agents, and cannot provide the sole basis of rational norms for human reasoners.

Theories of bounded rationality, which take into account the finite limitations of human cognitive ability, provide alternatives to accounts of perfect or idealized rationality which are based solely on abstract logical systems. According to bounded rationality, the prescriptiveness (or normative standing) of a set of norms derives not only from its connection to an abstract or formal standard (e.g. deductive closure, absence of contradiction, etc.) but also from the fact that such standards can be attained in principle by human reasoners.[26] That is, it is within our capacity to reason in accordance with the norms: that we ought to reason in a particular way implies that we can reason in that way.

### 2.2. Cognitive Biases in Reasoning

Despite the initial plausibility of the idea of bounded rationality, a question quickly emerges: to what degree should actual performance be taken as the measure of competence? Should the untutored behavior of normal cognitive agents serve as a basis for prescriptive theories of rationality? Problematically, there are many well known cognitive biases – fallacies of reasoning if you will – which typify human cognitive habits. Evans writes that "A 'bias' is usually defined as systematic attention to some logically irrelevant features of the task, or systematic neglect of a relevant feature."[27] One such characteristic failure of reasoning, which has come to be called the phenomenon of belief perseverance, was described by Francis Bacon in the *New Organon*:

> The human understanding when it has once adopted an opinion draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects, in order that by this great and pernicious predetermination the authority of its former conclusion may remain inviolate.[28]

It is this with cognitive bias that I am presently concerned, and I will consider it as it has been studied in a series of experiments known as the debriefing paradigm.

---

[26] Cf. Jonathan Baron, *Rationality and Intelligence* (Cambridge: Cambridge University Press, 1985) for a distinction between normative and prescriptive theories.

[27] Jonathan St. B.T. Evans, "Bias and Rationality," in *Rationality: Psychological and Philosophical Perspectives*, eds. Ken I. Manktelow and David E. Over (New York: Routledge, 1993), 16.

[28] Francis Bacon, *New Organon* (1620), quoted in Nisbett and Ross, *Human Inference*, 167.

## 3. The Debriefing Paradigm

### 3.1. The Paradigm Described

While familiar to psychologists, a recapitulation of the debriefing paradigm might still be in order. In general the debriefing paradigm works as follows: it is designed to prompt a participant reasoner (either an actor or an observer) to form a belief (e.g., their success at a given task) (actors form such beliefs about themselves, observers about actors), under controlled circumstances on the basis of certain evidence (feedback given during the performance of the task) which is subsequently undermined (the feedback is shown to be false); participants then report on the status of the resultant and related beliefs (e.g., actual success on given task, on future tasks, and in general on related tasks). The result is that beliefs so formed often survive the complete undermining of the evidence on the basis of which they were formed.

A standard experimental paradigm[29] involves getting participants to distinguish between supposedly authentic and fake suicide notes. During the task, participants are given false feedback which ranks their success as either above average, average, or below average. Following the task each participant is completely debriefed: the false and predetermined nature of the feedback is thoroughly explained. On the (standard) outcome debriefing condition, participants are told that the feedback was contrived and not in any way linked to their actual performance; they are shown the experimenter's instructions specifying the details of the feedback to be given and assigning them to the success, average or failure group.[30] Finally, as part of an ostensibly unrelated questionnaire, participants are asked to estimate their actual performance on the task they completed, and their prospects for future success both in similar tasks and in general.

The result is that even following debriefing, participants assigned to the success group (and their observers) ranked their abilities more highly than those assigned to the average or failing groups. Ross, Lepper and Hubbard summarized the results this way:

> even after debriefing procedures that led subjects to say that they understood the decisive invalidation of initial test results, the subjects continued to assess their performances and abilities as if these test results still possessed some validity.[31]

---

[29] Ross, Lepper, and Hubbard, "Perseverance in Self-Perception."

[30] cf. Ross, Lepper, and Hubbard, "Perseverance in Self-Perception," 883, 885; Cathy McFarland, Adeline Cheam, and Roger Buehler, "The Perseverance Effect in the Debriefing Paradigm: Replication and Extension," *Journal of Experimental Social Psychology* 43 (2007): 234, 235-6.

[31] Ross, Lepper, and Hubbard, "Perseverance in Self-Perception," 884.

That is, the inferred belief perseveres "[even] when a person discovers that the entire evidence base for the initial [inferred] judgement is not merely biased or tainted but is completely without value."[32] Despite the fact that most reasoners seem to behave in this way, the intuitive judgement of most theorists is that this behavior is irrational, and that reasoners ought to abandon, or at least revise their confidence in, the belief about their abilities based on the false feedback. Indeed, the results of the debriefing paradigm seem paradoxical because the reasoner, through the process of the experiment, recognizes the defeat of a belief (about the authenticity of the feedback) she has, but she does not recognize this as the defeat of a set of reasons on the basis of which she adopted an additional belief (about her ability).

Subsequent work with the debriefing paradigm[33] has not only confirmed the results of the initial studies, but has extended them in several directions. Perhaps most remarkable is a study by Wegner, Coulton and Wenzlaff[34] which demonstrated that even when the falsity of the feedback was made apparent to participants *prior* to their receiving it (i.e., briefing rather than debriefing), participants still treated the (false) information as though it were true and not completely undermined. "In sum," they wrote, "briefing and debriefing had essentially equivalent effects, leading neither actors nor observers to forsake the feedback as a cue to the actor's performance."[35]

Empirically, two debriefing techniques have been found to effectively the counter-act the perseverance effect. First, Ross, Lepper and Hubbard[36] reported that process debriefing, whereby in addition to a normal, outcome debriefing, participants are also informed about the phenomenon of belief perseverance and how it can occur, and then warned of its potential personal relevance to them in the context of the experiment, significantly reduced and often effectively eliminated the perseverance effect. Second, McFarland, Cheam and Buehler[37] reported that a revised outcome debriefing which, in addition to a normal outcome debriefing, informed participants of the invalidity of the entire test, eliminates the perseverance effect just as well as process debriefing. Revised outcome debriefing seeks to make manifest to the participant not merely that the information given in feedback is

[32] Nisbett and Ross, *Human Inference*, 176.

[33] e.g., McFarland, Cheam, and Buehler, "The Perseverance Effect in the Debriefing Paradigm."

[34] Daniel M. Wegner, Gary F. Coulton, and Richard Wenzlaff, "The Transparency of Denial: Briefing in the Debriefing Paradigm," *Journal of Personality and Social Psychology* 49 (1985): 338-346.

[35] Wegner, Coulton, and Wenzlaff, "The Transparency of Denial," 342.

[36] Ross, Lepper, and Hubbard, "Perseverance in Self-Perception."

[37] McFarland, Cheam, and Buehler, "The Perseverance Effect in the Debriefing Paradigm," 235-236.

false, but that the task itself is entirely fake, i.e., that the source of the information has no potential whatsoever to be reliable.[38]

### 3.2. Explaining our Intuitions Concerning the Rationality of Belief Perseverance

Perhaps the first step in appreciating the explanatory paradoxes raised by belief perseverance is to get a handle on the intuitions theorists tend to have concerning its irrationality. The reason behind this type of intuition seems to be something like this: wholly unjustified beliefs ought to be abandoned. As Ross, Lepper and Hubbard put it: "With no pertinent information remaining, ... the perceiver's assessment [should] return to its starting point as any logical or rational impression-formation model would demand."[39] Indeed, this intuition seems to be a corollary of some other, more general, principles of rationality.

The first of these might be the principle of *evidence proportionalism* which has been defined by Engel (following Hume[40]) as follows: "In general a belief is rational if it is proportioned to the degree of evidence that one has for its truth."[41] The rationality of belief is explained, at least in part, through a kind of evidence proportionalism. Thus, beliefs without any evidential support ought to be abandoned. Specifically, the persevering belief, having been deprived of the only supporting evidence on the basis of which it was adopted, seems to be completely unsupported, thereby rationally requiring its abandonment.

A second general principle which perhaps underlies our intuitions concerning the rationality of belief perseverance is the *principle of commutativity*, which Nisbett and Ross formulate as follows: "the net effect of evidence A followed by evidence B must [i.e., ought to] be the same as for evidence B followed by evidence A."[42] Roughly, the order in which we receive information ought not to affect its overall significance or evidential force. Bringing this principle to bear on belief perseverance treats it as a limiting case of the primacy effect whereby newly received information is synthesized in such a way as to represent is as consistent

---

[38] Seemingly, in Ross, Lepper, and Hubbard's initial (1975) experiment the task was genuine: there were authentic as well as inauthentic suicide notes in each pair, and thus a correct answer was possible on any trial. By the time McFarland, Cheam, and Buhler (2007) repeat the study, it seems that all the notes are inauthentic, and thus that the test itself was a fabrication, with no possibility of measuring what it purported to measure.

[39] Ross, Lepper, and Hubbard, "Perseverance in Self-Perception," 881.

[40] David Hume, *Enquiries Concerning Human Understanding and Concerning the Principles of Morals* (1777) (Oxford: Clarendon Press, 1975), X.i.87; 110.

[41] Pascal Engel, "Introduction: The Varieties of Belief and Acceptance," in *Believing and Accepting*, ed. Pascal Engel (Dordrecht: Kluwer, 2000), 3.

[42] Nisbett and Ross, *Human Inference*, 169.

and coherent with previously received information. That is, reasoners are maximally conservative when revising existing beliefs, tending instead to interpret the significance of new information in light of already accepted information. Recognizing that the order in which we receive information is normally irrelevant to its evidential import, the principal of commutativity condemns the perseverance effect as irrational.

### 3.3. Explaining the Results of the Debriefing Paradigm

A variety of theories exist that contribute to a psychological explanation of the phenomenon of belief perseverance. Anderson[43] discusses three psychological processes that contribute to such an explanation: the availability heuristic (where only memorable confirming or disconfirming cases are considered); illusory correlation (where more confirming cases and fewer disconfirming ones are remembered than actually exist); and data distortions (where "confirming cases are inadvertently created and disconfirming cases are ignored").

The operation of these cognitive biases can easily be appreciated when imagining the reasoning process of participants in the debriefing paradigm. Nisbett and Ross imagine participants, having adopted the target belief, searching around for confirming evidence among their existing beliefs. For example, a participant who is given (false) feedback that she is a successful discriminator of genuine versus faked suicide notes could take her "reasonably good performance in her abnormal psychology course, her ability to make new friends easily, and her increasing sense of confidence and assurance as she progressed in the … task" as "further evidence" of her discriminatory powers. Similarly, a participant given negative feedback "might note her difficulty in imagining herself as lonesome or alienated, her mediocre performance in her social problems course, and her increasing sense of confusion and hesitation as she progressed in the … task" as further evidence of her own unreliability.[44] There is adequate confirming evidence no matter which way things go. This scenario can occasion an important hypothesis, for it promises to show something important about how humans learn and synthesize information. When we acquire a new piece of information we synthesize it with our existing beliefs by finding ways that it can serve as a premise or conclusion from our existing beliefs. As we find new ways that the belief can be a conclusion of our existing beliefs, it becomes further entrenched in our overall web of belief.

[43] Anderson, "Belief Perseverance," 110.

[44] Nisbett and Ross, *Human Inference*, 181.

Another explanatory hypothesis, first suggested by Ross, Lepper and Hubbard[45] and later by Anderson, Lepper and Ross[46] is that the participant finds *explanations* of the result from among their existing beliefs. That is, the new piece of information is treated as an explanandum, and reasoners search around for an explanans among their existing beliefs. Here, the participant's beliefs about her performance in her abnormal psychology or social problems course does not serve as evidence for her belief about her 'performance' (as described by the false feedback) in the experimental task, but rather serves to causally explain her 'performance' in the experiment. Since the explanans already occurs among her existing beliefs (i.e., the usual evidential order for a causal explanation is reversed), this further allows the participant to accept the belief because, having explained it, they are entitled to expect the result, or better understand why it happened. Problematically, though, as Anderson, Lepper and Ross observe, "[o]nce a causal account has been generated, it will continue to imply the likelihood of the 'explained' state of affairs even after the original basis for believing in that state of affairs has been eliminated."[47]

Evolutionary explanations tend to claim that the acquisition of new information is costly. Therefore, once a piece of information has been acquired a cognitively economical strategy is to be as conservative as possible when it comes to revising or abandoning one's doxastic attitude to that information.

Having surveyed some of the explanations of the phenomenon of belief perseverance arising from the debriefing paradigm, the task of more closely scrutinizing the rationality of the behavior remains.

## 4. Foundations and Coherence

In *Change in View* Harman examines the relationship between justification and belief revision, considering two theories of justification: the foundations theory and the coherence theory. As a theory of justification, each theory serves as a model of ideal belief revision.[48] In distinguishing the two theories, Harman writes: "[t]he key issue is whether one needs to keep track of one's original justifications for beliefs," foundationalists say "yes" and coherentists say "no."[49] Thus, Harman writes, "the theories are most easily distinguished by the conflicting advice they occasionally give concerning whether one should *give up* a belief P ... when P's original

[45] Ross, Lepper, and Hubbard, "Perseverance in Self-Perception," 890.
[46] Anderson, Lepper, and Ross, "Perseverance of Social Theories."
[47] Anderson, Lepper, and Ross, "Perseverance of Social Theories," 1038.
[48] Harman, *Change in View*, 29.
[49] Harman, *Change in View*, 29.

justification has to be abandoned."[50] As theories of justified belief, foundations and coherence can be roughly characterized as follows:

- *Foundations theory of justified belief*: one is justified in continuing to believe something only if one has a special reason to continue to accept that belief;
- *Coherence theory of justified belief*: one is justified in continuing to believe something as long as one has no special reason to stop believing it.[51]

Corresponding to these two theories of justified belief, are two corollary positions concerning belief revision.[52] Deriving from the foundations theory is the

> *Principle of negative undermining*: One should stop believing *P* whenever one does not associate one's belief in *P* with an adequate justification (either intrinsic or extrinsic),[53]

while the coherence theory gives us the

> *Principle of positive undermining*: One should stop believing *P* whenever one positively believes one's reasons for believing *P* are no good.[54]

As Harman construes it, the foundations theory holds that instances of belief perseverance arising from the debriefing paradigm violate the principle of negative undermining and are therefore irrational, while the coherence theory licenses this behavior as being consistent with the principle of positive undermining and therefore rational.

As Harman observes, when we consider actual cases of belief perseverance it becomes clear that there are problems with the descriptive accuracy of the foundations theory. It is found that people retain beliefs even after the positive refutation of all the evidence that was originally provided in support of the belief. To explain this phenomenon, Harman suggests that

---

[50] Harman, *Change in View*, 30.

[51] Harman, *Change in View*, 36.

[52] Harman, *Change in View*, 39.

[53] For Harman (*Change in View*, 30-31), the justification of basic beliefs is intrinsic while derived beliefs – beliefs whose justification relies on other beliefs – have extrinsic justifications.

[54] In considering the debriefing paradigm, Goldman proposes a similar rule. Having rejected a rule which instructs one to "Revise all ...beliefs that have been undermined by new evidence," and even a "rule system [which] would *oblige* a cognizer *continually* to search for old beliefs in LTM [long term memory] that might be weeded out in light of new evidence," he proposes a rule which prescribes "if one activates an old belief in q, and if one (actively) believes that this belief wholly stems from now abandoned evidence, then one is required to abandon q" (Alvin I. Goldman, *Epistemology and Cognition* (Cambridge: Harvard University Press, 1986), 220-21).

> what the debriefing studies show is that people simply do not keep track of the justification relations among their beliefs. They continue to believe things after the evidence for them has been discredited because they do not realize what they are doing. They do not understand that the discredited evidence was the *sole* reason why they believe as they do.[55]

Yet, Harman argues that the foundations theory not only fails descriptively; it also fails as a prescriptive theory. Harman argues that "[P1] the [foundations] theory implies that people are unjustified in almost all their beliefs. [And P2] This is an absurd result." The reasons Harman gives for P1 are [Pi] that the foundations theory requires that agents keep track of their reasons for their beliefs in order for their beliefs to be rationally justified and, [Pii] as evidenced by studies like the debriefing paradigm, "people rarely keep track of their reasons [for their beliefs]."[56] Indeed the computational costs for any cognitively finite agent attempting to actively track all of the evidentiary relations on the basis of which it (comes to) hold(s) its beliefs is intractable.[57] Since any normative theory which classifies the majority of our beliefs as irrational violates Cohen's basic principle of judging instances of irrational performance against a backcloth of rational competence, the foundations theory cannot be accepted as providing the normative standards against which the rationality of our inferential behavior should be assessed. Therefore, Harman concludes, "The foundations theory turns out not to be a plausible normative [i.e., prescriptive] theory [of rational belief change] after all."[58]

By our rational intuitions alone, we tend to judge the belief perseverance behavior of participants in the debriefing paradigm as irrational. Yet according to Harman, these rational intuitions presuppose a foundations approach to reasoned belief revision which is unacceptable. Further, since on the coherence theory the majority of our actual beliefs are rationally held, it is a better prescriptive theory than foundationalism. As a corollary, belief perseverance behavior also turns out to

---

[55] Harman, *Change in View*, 38. This seems to be a point which Ross, Lepper and Hubbard themselves accepted. They wrote: "We propose that first impressions may not only be enhanced by subsequent biases in coding but may ultimately be *sustained* through such biases. The perceiver, we contend, typically does not reinterpret or reattribute impression-relevant data when the basis for his original coding bias is discredited; *once coded, the evidence becomes autonomous from the coding scheme, and its impact ceases to depend upon the validity of that schema*" (Ross, Lepper, and Hubbard, "Perseverance in Self-Perception," 889, emphasis added).

[56] Harman, *Change in View*, 39.

[57] Mike Oaksford and Nick Chater, "Reasoning Theories and Bounded Rationality," in *Rationality: Psychological and Philosophical Perspectives*, 31-60; Mike Oaksford and Nick Chater, "Theories of Reasoning and the Computational Explanation of Everyday Inference," *Thinking and Reasoning* 1 (1995): 121-152.

[58] Harman, *Change in View*, 39.

be rational. Basically, Harman's view (as I read it) is that our intuitions about the rationality of belief perseverance and the debriefing paradigm are irrational not the behavior of reasoners in these cases.

## 5. Debriefing: Why Failure to Track Reasons is not the Problem

Against Harman, I hold that our intuitions about the irrationality of belief perseverance in cases like the debriefing paradigm are correct. In responding to his position I do not propose an argument in favor of the foundations theory of rational belief revision. Rather, I deny that our intuitive judgements of the irrationality of belief perseverance presuppose the foundations theory. Instead, I provide an account of what is going on in the debriefing paradigm which confirms our intuitions about the irrationality of the results of the paradigm without requiring that reasoners track their reasons for their beliefs.

On Harman's account, the failure of the reasoner is not one of rationality but one of memory. She has simply forgotten that the defeated reasons were the reasons on the basis of which she initially adopted a belief and it is because of this that she does not draw the connection between the defeat of those reasons and the acceptability of the belief. On this picture, even the failure to actively track her reasons – regardless of whether those reasons become defeated or overridden – is enough to violate the principle of negative undermining thereby requiring abandoning the belief. By contrast, the principle of positive undermining is never violated since the reasoner never realizes that her *reasons* have been defeated. That is, she never comes to the positive belief that her reasons for her belief are no good: not because she does not recognize the defeat of certain claims (the reasons themselves) but because she fails to connect those claims with the belief for which they served as reasons.

But the question remains: is this failure to make the connection between the defeat of the reasons and the acceptability of the belief properly explained as a failure of memory? I maintain that the results of the debriefing paradigm, and others like it, are not cases of negative undermining and are not properly explained or justified as a failure of memory. Instead, these are cases of positive undermining, where reasoners fail to recognize the evidentiary significance of new information available to them. Even if the reasoner has not tracked (e.g., by forgetting) her initial reasons for adopting the belief, this neither explains nor excuses her failure to examine the acceptability of the belief following the defeat of those reasons. On my account, the failure in such cases is not an understandable failure of memory but a rationally reprehensible failure to see the immediate consequences of new information. This failure to recognize the immediate consequences of new information is best understood not as a failure of memory but as a failure of understanding.

It is important to note that accounts involving both the foundations theory and the coherence theory presuppose that the reasoner has not arrived at any *new* or *alternate* reasons for the target belief. If it is supposed that the reasoner possesses any new or alternate reasons supporting the belief, then the case ceases to be problematic: we would not intuitively deem it irrational that she would continue to hold the belief, and neither would either of the two theories.

### 5.1. Understanding and Drawing the Right Inferences

In the debriefing paradigm, the participant reasoner drew an inference regarding her abilities on the basis of the information given to her as feedback during the experiment. We might call the connection in her mind between the feedback and the conclusion she drew therefrom her cognitive warrant.[59] That is to say that part of the significance to the reasoner of the information given in the feedback is that it yielded certain consequences. Yet when, during the debriefing, the reasoner recognized that this same information was defeated, she did not thereby recognize that the conclusion she drew therefrom might be undermined. That is, she failed to appreciate the immediate consequences of new information she had come to accept. There is no need for her to have tracked or remembered the initial inference she drew. Rather, the only thing that is required of her is that she recognize the significance of the information immediately before her. On this account the failure is not one of memory but of understanding.

Part of what it is to understand a claim is to understand how it connects inferentially to other claims. That is, part of what it is to understand a claim is to know what other claims it could be concluded from, and what other claims it could serve as a premise for. This view of meaning and understanding Brandom calls *inferentialism*[60]

> Understanding or grasping a propositional content is here presented not as the turning on of a Cartesian light, but as a practical mastery of a certain kind of inferentially articulated doing: responding differentially according to the circumstances of proper application of a concept, and distinguishing the proper inferential consequences of such application. … Thinking clearly is on this inferentialist rendering a matter of knowing what one is committing oneself to by a certain claim, and what would entitle one to that commitment. … Failure to

[59] See the next section for an explanation of the notion of a cognitive warrant.

[60] With Brandom, I see this as part of a pragmatic view of meaning which holds that meaning is explained in terms of use, and to understand a linguistic expression is to know how it is correctly used. One of the ways that we use statements is as premises and conclusions in inferences. Thus, part of what it is to understand a claim is to know what inferences can be correctly made involving them.

> grasp either of these components is failure to grasp the inferential commitment that use of the concept involves, and so failure to grasp its conceptual content.[61]

While Brandom here talks in terms of understanding propositions, elsewhere he specifically links this inferentialist account of understanding to the contents of beliefs, writing: "Understanding the content of a speech act or belief is being able to accord the performance of that speech act or the acquisition of that belief the proper practical significance – knowing how it would change the score [of commitments and entitlements] in various contexts."[62] So, to correctly understand the meaning of a claim is to understand its inferential significance. To the degree to which we fail to appreciate the consequences of a claim (the commitments it puts upon us), we fail to understand it. Similarly, to the degree to which we fail to appreciate what a claim is a consequence of (what would entitle us to it), we fail to understand it.

In the case of the debriefing paradigm, the participant reasoner fails to appreciate the commitments put upon her by her accepting during debriefing that the information given as feedback is indeed false. What makes cases like this so remarkable is not that an individual has forgotten her reasons for adopting a belief. Rather, it is that on one occasion she recognized some information as having a certain significance by immediately making certain inferences on its basis. Yet, on another occasion she fails to recognize the significance of that same information by failing to make the relevant inferences. Seen in this way, the reasoner fails to appreciate the meaning of the information, or at least treats the information differently from one occasion to the next. Rather than forgetting their reasons for a belief which they may not be attending to, reasoners in the debriefing paradigm fail to appreciate the significance of information immediately present to them and thereby misunderstand it.

---

[61] Robert B. Brandom, *Articulating Reasons: An Introduction to Inferentialism* (Cambridge: Harvard University Press, 2000), 63-64.

[62] Brandom, *Articulating Reasons*, 165-166. The passage preceding the quoted sentence reads as follows: "One can pick out what is *propositionally* contentful to begin with as whatever can serve both as a premise and as a conclusion in *inference* – what can be offered as, and itself stand in need of, *reasons*. Understanding or grasping such a propositional content is a kind of know-how – practical mastery of the game of giving and asking for reasons, being able to tell what is a reason for what, distinguish good reasons from bad. To play such a game is to keep *score* on what various interlocutors are committed and entitled to. Understanding the content of a speech act or belief is being able to accord the performance of that speech act or the acquisition of that belief the proper practical significance – knowing how it would change the score in various contexts" (Brandom, *Articulating Reasons*, 165-166).

Two situational facts about the experimental method of the debriefing paradigm make this result especially conspicuous. First is the fact that the inferential path from the feedback to the target belief was not complicated but was an immediate inference for the participant reasoner. No other information was required in making the inference in the initial instance. Second is the fact that the time between the initial feedback and its subsequent undermining is not especially long.[63] *Ex hypothesi*, not only was there no occasion for the discovery of *new* evidence for the target belief but there was no occasion for the re-evaluation or displacement of the cognitive warrant initially relied upon.

Viewed in this way, the prescribed result of the debriefing paradigm does not seem nearly as cognitively onerous as Harman's account makes it out to be. Yet the question remains as to whether the situation of the debriefing paradigm is properly interpreted as an instance of (unrecognized) positive undermining. To answer this question we must consider the notion of a cognitive warrant a little more closely.

### 5.2. Cognitive Warrants and Positive Undermining

Harman's *principle of positive undermining* states: "one should stop believing *P* whenever one positively believes one's reasons for believing *P* are no good."[64] The question is, how are external judges to determine when an individual reasoner judges – or ought to judge – that her reasons for believing something are good or no good? To make this determination solely on the basis of whether the reasoner actually comes to adopt or abandon some belief cannot be accepted as a satisfactory. The problem with this method is that it presumes that the reasoner is never mistaken or irrational. Yet since cases like belief perseverance raise questions about the rationality of individual reasoners in certain situations, we should not allow this question to be begged.

Instead, what is needed is a way of determining when the principle of positive undermining has been satisfied, even if a reasoner has failed to recognize this or to act on it appropriately. I suggest that a neutral way of attempting to determine when a reasoner (ought to) positively believe(s) that her reasons for believing something are (no) good is to invoke the idea of a cognitive warrant.

---

[63] Ross, Lepper, and Hubbard ("Perseverance in Self-Perception") initially tested for two intervals, 5-minutes and 25-minutes between the conclusion of the briefing-phase and debriefing. (During this time participants are not exposed to any new information.) Finding no statistically significant difference between these conditions, they opted for the shorter interval. Similarly, McFarland, Cheam and Buehler use a delay interval of "a few minutes" (McFarland, Cheam, and Buehler, "The Perseverance Effect in the Debriefing Paradigm," 235).

[64] Harman, *Change in View*, 39.

A warrant is like an inference ticket: it is a rule that licenses or underwrites a move to infer some claim on the basis of other claims. A cognitive warrant is the warrant that a reasoner actually uses or actually draws upon in making an inference on some particular occasion. Cognitive warrants can be understood as something like the "habits of mind" Peirce described in "The Fixation of Belief" as "that which determines us, from given premises, to draw one inference rather than another" and which Peirce there called a *guiding principle* of inference.[65]

A cognitive warrant is psychological in nature and need not be explicitly formulated in the mind of the reasoner in order to be operative. A cognitive warrant might have no logical or epistemological merit whatsoever instead being based on nothing more than a psychological association in the mind of a reasoner. But even as such, it has psychological force for that individual reasoner. Further, a reasoner need not articulate the warrant to herself when relying upon it in her reasoning; indeed she need not be aware of it at all. Cognitive warrants are markers of consequences that some reasoner finds immediately apparent when presented with certain information. It is the job of epistemologists and psychologists of reasoning to make these cognitive warrants explicit in an attempt to explain and evaluate processes of reasoning.

Importantly, Harman's account of reasoning relies on notions very similar to these cognitive warrants or guiding principles of inference. Harman assumes that "one has certain basic dispositions to take some propositions immediately to imply other propositions and to take some propositions as immediately inconsistent with each other,"[66] thereby introducing the notions of immediate implication and immediate inconsistency which are defined internally to the psychology of individual reasoners. These notions are meant to replace the purely logical notions of implication and inconsistency in formulating prescriptive norms for reasoners. For example, Harman's *Principle of Immediate Implication* states: "That *P* is immediately implied by things one believes can be a reason to believe *P*."[67] In the debriefing paradigm, it would seem that, even on Harman's account, the target belief concerning a reasoner's task-related abilities is an immediate implication of the feedback given during the experiment. Whatever the link is in the mind of the reasoner between the feedback and the target belief we can call her cognitive warrant.

[65] Peirce, "The Fixation of Belief."
[66] Harman, *Change in View*, 19.
[67] Harman, *Change in View*, 21.

The truth of such guiding principles of inference, according to Peirce, "depends on the validity of the inferences which the habit determines."[68] Thus, cognitive warrants can be objectively evaluated against some external standard of rationality according to whether they are (generally) truth-preserving or reliable. But, more importantly for the purposes of this argument, the reasoning of individuals can also be rationally evaluated by an internal standard according to whether the agent consistently applies the cognitive warrants they actually (though perhaps tacitly) accept from one occasion to the next.

That a reasoner relies on, or acts in accordance with, a cognitive warrant on some occasion is an indication that she finds it to mark a relationship of immediate implication between two (sets of) beliefs. That is, at some (perhaps unconscious) level she finds the premissiory beliefs to be good reasons for the conclusion-belief. Because of this, we should be able to use the notion of cognitive warrants to determine when the *principle of positive undermining* ought to apply. Namely, we can say that a reasoner, S, ought to believe that her reasons for believing that *P* are no good whenever a set of beliefs that immediately imply *P* (for S) have been manifestly (to S) shown to be unacceptable. That is, a belief, *P*, has been positively undermined whenever a cognitive warrant having *P* as its conclusion has been positively undermined. Yet, this is exactly what happens in the case of the debriefing paradigm: participants are shown that their reasons – reasons which immediately implied some target belief – are no good.

To better appreciate this, consider the way that Ross, Lepper and Hubbard described standard outcome debriefing:

> The experimenter explained that the subject's success or failure had been randomly determined prior to her arrival. He emphasized that the subject's score had not been dependent on her performance and that it provided absolutely no information about her actual performance.[69]

Similarly, McFarland, Cheam and Buehler describe (standard) outcome debriefing as follows:

> Participants in the *standard outcome debriefing* condition were informed that their score was a fake score that had been randomly assigned to them prior to their arrival. Additionally, they were shown a "random assignment schedule," and the experimenter emphasized that the score contained absolutely no information about the participant's actual performance or underlying abilities.[70]

---

[68] Peirce, "The Fixation of Belief."

[69] Ross, Lepper, and Hubbard "Perseverance in Self-Perception," 885.

[70] McFarland, Cheam, and Buehler, "The Perseverance Effect in the Debriefing Paradigm," 235.

The emphasis offered in debriefing has the effect of reminding the participant of the inference she had drawn only minutes ago on the basis of the now defeated information. In doing so, it makes manifest to the participant that the cognitive warrant she relied on just previously is entirely defeated and her inference thereby undermined.

Seen in this way, the debriefing paradigm is a situation of positive rather than negative undermining. Recognizing that the undermining occurs does not require tracking any reasons, but only requires that the reasoner recognizes the immediate implications of the information presented, indeed emphasized, to her in debriefing. Her failure to see that the target belief has been undermined is evidence not that she is forgetting what her reasons for that belief were, but that she is misunderstanding (or ignoring the probative significance of) the information immediately before her by failing to apply the same cognitive warrant on one occasion that she had (and was reminded she had) applied only minutes previously. And, in accordance with our intuitions, this behavior is irrational.

The problem, I suggest, with Harman's approach to assessing the rationality of belief perseverance in debriefing is not his advocacy of a coherentist principle of positive undermining over a foundationalist principle of negative undermining. Rather, the problem is that he psychologizes the criteria for determining when positive undermining has occurred. Positive undermining occurs when one comes to believe that one's reasons for believing some claim are no good. Yet, as a criteria for determining when this occurs, Harman seems to invoke his *Immediate Inconsistency Principle*: "Immediate logical inconsistency in one's view can be a reason to modify one's view."[71] It would seem that Harman takes positive undermining to have occurred only if an agent positively forms a belief that the defeat of her reasons (during debriefing) is inconsistent with her acceptance of the target belief. In other words, positive undermining has not occurred unless the reasoner *recognizes* not merely that her *reasons* for P are defeated, but also that this is somehow inconsistent with continuing to hold that P. But to use this as the criterion for positive undermining is to presuppose that the reasoner is always rational, and it is precisely the rationality of the reasoner that is at issue. Positive undermining can occur, and yet the reasoner can fail to recognize it.

This problem is overcome, I suggest, by the proposed method. The cognitive warrant method is neutral concerning objectively correct rational norms because it does not impose any external normative standard of reasoning on the reasoner, instead relying on her own inferential habits to determine which inferences she takes to be good ones. At the same time, it allows third-party judges to hold

[71] Harman, *Change in View*, 22.

reasoners accountable to their own putative standards, by insisting that they give the same information the same significance from one occasion to the next by applying the same cognitive warrants to it as they have in the immediate past. On this model it is possible for a reasoner to fail to recognize that a belief has been positively undermined, and it is possible for a third-party to make a judgement about when this has occurred.

This is not to say that reasoners cannot abandon or change their conscious attitudes towards their cognitive warrants just as they can their beliefs. It does, though, allow theorists of reasoning a way of making explicit those rules which characterize a reasoner's inferential habits, and thereby of talking about when it is rational for reasoners to change those inferential habits. Because of the experimental conditions of the debriefing paradigm and the temporal proximity from feedback to debriefing it is not reasonable to suppose that the cognitive warrant relied upon in the feedback stage should have been abandoned or re-evaluated prior to debriefing.

### 5.3. An Objection to the Proposed Account

An important objection to the account just described is that the defeat of antecedent information occasioning the drawing of a specific inference need not undermine the conclusion drawn in that inference.

Let us represent the cognitive warrant used by the participant reasoner, S, as 'R → C', such that, in accepting R, S takes C to be immediately implied and is thereby cognitively compelled to infer C. Importantly, there is nothing in this cognitive warrant that compels S to conclude ~C on the basis of ~R. That is to say '~R → ~C' need not be a cognitive warrant of S. Indeed, ~R and C may be entirely consistent not only with each other but with the remainder of S's beliefs also. More importantly, ~R and C may not seem immediately inconsistent to S and this seems to challenge my interpretation that the defeat of R is an instance of positive undermining.

Why should S's subsequent acceptance of ~R have any effect whatsoever on her attitude towards C? And what effect should that be?

The answer to the first question is that positive undermining involves forming the positive belief that one's reasons for a belief are no good, not in believing there to be an immediate inconsistency between two claims. The status of R as one of S's reasons for C is not established by S's remembering the inferences she made on the basis of R, but by the significance S takes R to have. That S takes R to be a reason for C is shown by the fact that she took C to be immediately implied by R. There is a cognitive connection in S's mind between R and C. Because of this

cognitive connection, the recognition of R's defeat should call to mind the acceptability of C. That C is immediately implied by R (for S) is enough to say that the condition of positive undermining has been met when R is recognized as false by S. Because of this, the undermining of R should have some effect on the cognitive attitude or a reasoner towards C.

Should it require the abandoning of C? No – so long as S has other adequate reasons justifying the belief. (Recall that the normative debate concerning the rationality of belief perseverance in debriefing presupposes that this does not occur and each side agrees on the rationality of the result if it does occur.) What the undermining of a reason should do is call into question the acceptability of the belief thereby altering one's cognitive attitude towards it. In the absence of other, immediately apparent evidence for C, the defeat of R should reduce S's confidence in C. When C is a matter of pressing importance for S (e.g., it is a matter which needs to be settled right away) the defeat of R should also occasion the search for other reasons for C. In some circumstances (e.g., depending on what is at stake) the defeat of R should decrease if not eliminate S's reliance on C, (e.g., when selecting premises for subsequent reasoning). Should S find that all of her reasons for believing C are no good – i.e., that she has no good reason whatsoever for believing that C – then she should be rationally obliged to abandon C altogether. So, while the defeat of R need not require the abandoning of C, it is a case of positive undermining, and to suggest, as the coherence theory does, that no change in S's cognitive attitude is required cannot be taken to be good rational advice.

In summary, the account I have proposed is like Harman's coherence theory in that it does not require that rational agents track their reasons; it "does not suppose there are continuing links of justification dependency that can be consulted when revising one's beliefs."[72] Rather, it claims that part of the significance information has for reasoners is the role it plays in the inferences they make. Further, failure to make the right sorts of inferences with a given piece of information is a failure to understand the significance of that piece of information. The proposed account is consistent with existing psychological explanations of the perseverance effect without sanctioning the behavior is rational. It is perhaps quite understandable that a reasoner might mistake the overall (evidentiary and explanatory) coherence of her beliefs as evidence for the acceptability of some particular belief, C, without realizing that the remainder of her beliefs would cohere equally with C's opposite. Yet, in the absence of a reason to accept C instead of ~C, one's cognitive attitude towards each should not be radically different.

[72] Harman, *Change in View*, 39.

## 6. Bounded Rationality Revisited

Intuitively it would seem that any principles of rationality that are to guide or regulate our thinking must be principles that we are capable, at least in principle, of following. But this does not mean that our everyday performance has to be sanctioned as rational.

There are several problems with the unqualified claim that principles of rationality must be competence norms. In the first place, how is competence to be determined? Perhaps the day-to-day performance of untutored reasoners is not the best measure of competence. Especially if the habits of mind which serve as the guiding principles of the inferences we make can be altered with training – as is the hope of every course in reasoning skills and critical thinking. The point of teaching and learning reasoning skills is to train the mind to habitually invoke or rely upon good cognitive warrants and to detract from the reliance upon guiding principles of inference that are unsound. Moreover, in judging some of our performances to be rational and others of them not so, we appeal to some standard or ideal which, while we can grasp, we do not always attain. Thus there must be some measure of rationality beyond our own behavior or cognitive habits to which we appeal when conceptualizing rationality.

Second, even if we accept that theorists must presume a background of rational competence against which the performance of individual acts of reasoning are measured, this need not require theorists to presume that all human cognitive habits, strategies and tendencies are rational. It is entirely possible that generally competent human reasoners have a variety of cognitive tendencies which, while wholly or generally unreliable, they nevertheless rely upon with predictable regularity. To suppose that we are rationally competent in general does not mean that there are not systematic ways in which we fail to be rational. Many of these cognitive biases are well-known and have been widely studied. Contrary to Harman's coherence theory,[73] a belief does not acquire justification simply by being believed. Rather, if some of our ways of acquiring or preserving beliefs are not always rational, then the mere fact of believing does on its own count as a reason for the justifiability of the belief, let alone show that belief to be justified. Instead, it must additionally be shown that the believer is being rational in believing what she does. More generally, our descriptions and theories of human reasoning behavior should not exclude the very possibility that some failure of rationality can occur on any particular occasion, even if we accept on principle that it cannot occur in every occasion or even on most occasions.

[73] Harman, *Change in View*, 35.

Finally to challenge the idea that constitutive competencies should serve as the final ground for our normative ideals of rationality, consider the following case. Imagine a group of reasoners who, for whatever reason, are *by our lights* constitutionally incompetent in a certain respect. Perhaps they perennially draw inferences that lead them into self-deception or akrasia despite our best pedagogical efforts and attempts to bring this to their attention. What are we to say of such a group? Should it be said that this group is perfectly rational in their own way, according to their own competence level? Instead, might we not want, while recognizing the cognitive limitations of such a group, to be able to say that they are irrational in certain specifiable ways? Indeed, suppose that there are those among us who *do* track our reasons in certain types of situations, perhaps by taking a mental note of them. Are we then to say that there are two standards of rationality, one for people with good memories and another for those with poor memories? Conceiving of rational norms solely as competence norms, combined with the view that there are different levels of rational competency, leads to the problem of relativism about the norms themselves. To say that our own competencies are the final ground for our rational norms is to say that it is inconceivable that we are somehow constitutionally irrational in certain respects. Yet, as the above example shows, there is no inconsistency in supposing this. Rather, we are rational to the extent that we are capable (constitutionally or otherwise) of acting in accordance with a set of standards and principles which are external to us. If we are unable to live up to those standards because of our cognitive constitutions we may not be faulted for this, but that does not make the behavior rational.

In the end, perhaps there are two morals to this story. First, one way to improve our overall rationality by is lowering our standards and expectations. And second, even proposing competence norms of which we feel ourselves capable involves picturing an ideally competent reasoner, or a set of standards against which our competence can be measured.