Current Issues in Thinking and Reasoning
Series Editor: Linden Ball

Current Issues in Thinking and Reasoning is a series of edited books which will reflect the state of the art in areas of current and emerging interest in the psychological study of thinking processes.

Each volume will be tightly focussed on a particular topic and will consist of from seven to ten chapters contributed by international experts. The editors of individual volumes will be leading figures in their areas and will provide an introductory overview.

Example topics include thinking and working memory, visual imagery in problem solving, evolutionary approaches to thinking, cognitive processes in planning, creative thinking, decision making processes, pathologies of thinking, individual differences, neuropsychological approaches and applications of thinking research.

Emotion and Reasoning
Edited by Isabelle Blanchette

New Approaches in Reasoning Research
Edited by Wim De Neys and Magda Osman

The Developmental Psychology of Reasoning and Decision-Making
Edited by Henry Markovits

Aberrant Beliefs and Reasoning
Edited by Niall Galbraith

Reasoning as Memory
Edited by Aidan Feeney and Valerie A. Thompson

Individual Differences in Judgement and Decision Making
Edited by Maggie E. Toplak and Joshua Weller

Moral Inferences
Edited by Jean-François Bonnefon and Bastien Trémolière

MORAL INFERENCES

Edited by
Jean-François Bonnefon and Bastien Trémolière

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2017
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book has been requested

ISBN: 978-1-138-93797-0 (hbk)
ISBN: 978-1-138-93798-7 (pbk)
ISBN: 978-1-315-67599-2 (ebk)

Typeset in Bembo
by Swales & Willis Ltd, Exeter, Devon, UK

CONTENTS

10

RATIONALIZATION IN MORAL AND PHILOSOPHICAL THOUGHT

Eric Schwitzgebel and Jonathan Ellis[1]

Abstract

Rationalization, in our intended sense of the term, occurs when a person favors a particular conclusion as a result of some factor (such as self-interest) that is of little justificatory epistemic relevance, if that factor then biases the person's subsequent search for, and assessment of, potential justifications for the conclusion. Empirical evidence suggests that rationalization is common in people's moral and philosophical thought. We argue that it is likely that the moral and philosophical thought of philosophers and moral psychologists is also pervaded by rationalization. Moreover, although rationalization has some benefits, overall it would be epistemically better if the moral and philosophical reasoning of people, including professional academics, were not as heavily influenced by rationalization as it likely is. We discuss the significance of our arguments for cognitive management and epistemic responsibility.

Introduction

People often seek justifications for conclusions they already believe or ideas they strongly favor. In some cases, what results is *rationalization*. Rationalization, as we will understand it in this chapter, occurs when a person favors a particular conclusion as a result of some factor (such as self-interest) that is of little justificatory epistemic relevance. The thinker then seeks an adequate justification for that conclusion, but the very factor responsible for her favoring it now biases how this search for justification unfolds. As a result of an epistemically illegitimate investigation, the person identifies and endorses a justification that makes no mention of the distorting factor that has helped guide her search.

In the second section, we will expand on this characterization of rationalization. But first consider two cases:

Newspaper. At the newsstand, the man selling papers accidentally gives Dana[2] a $20 bill in change instead of a $1 bill. Dana notices the error right away. Her first reaction is to think she got lucky and doesn't need to point out the error. She thinks to herself, "What a fool! If he can't hand out correct change, he shouldn't be selling newspapers." Walking away, she thinks, "And anyway, a couple of times last week when I got a newspaper from him it was wet. I've been overpaying for his product, so this turnabout is fair. Plus, I'm sure almost everyone just keeps incorrect change when it's in their favor. That's just the way the game works." If Dana had seen someone else receive incorrect change, she would not have reasoned in this way. She would have thought it plainly wrong for the person to keep it.

Kant-Objector. Kant's *Critique of Pure Reason*—a famously difficult text—has been assigned for a graduate seminar in philosophy. Mitchell, a student in that seminar, loathes Kant's opaque writing style and the authoritarian tone he thinks he detects in Kant. He doesn't fully understand the text or the critical literature on it. But the first critical treatment that he happens upon is harsh, condemning most of the central arguments in the text. Because he detests Kant's writing style, and without much consideration of possible Kantian counterarguments, Mitchell immediately embraces that critical treatment, and now he deploys it to justify his rejection of Kant's arguments. He would happily abandon that critique in favor of a different set of harsh critiques but he does not consider more charitable approaches.

The human capacity for rationalization has long been noted by playwrights, satirists, and philosophers, especially when it comes to matters of morality, obligation, and transgression, as in Newspaper. Early clinical psychologists also made it a point of emphasis (e.g., Freud, 1911; Jones, 1908). More recently, bias and distortion in moral reasoning have drawn the attention of cognitive scientists. What, though, is the prevalence and role of rationalization in the reasoning of philosophers, ethicists, and cognitive scientists themselves? What impact does rationalization have upon *their* reasoning in their own philosophical and moral thinking?

We will argue that both empirical psychology and philosophical reflection suggest that rationalization might play a large role in the professional work of philosophers and moral psychologists. We will also argue that the epistemic costs of rationalization are formidable: Rationalization typically results in epistemically unwarranted degrees of confidence, if not false belief; it obstructs the critical evaluation of one's own reasoning; and it impedes the productive exchange of reasons and ideas among well-meaning interlocutors.

Rationalization characterized

In the philosophical literature, rationalization is typically characterized as involving a difference between the justifications one offers in defense of an action or attitude and what really explains one's action or attitude (Audi, 1985; Siegel, 2014; Summers, manuscript).[3] We accept this broad characterization, with some clarifications and caveats.

First, we take the *target* of rationalization to be a belief or belief-like attitude toward some proposition P. (A "belief-like" attitude could be a judgment, an

acceptance, an expressed opinion, etc. For simplicity we will speak only of belief.) Rationalizing an action can then be understood as rationalizing a certain *belief about* that action, such as that it was morally permissible or not foolish. Similarly for the rationalization of desires and other non-belief-like attitudes.

In rationalization, one offers one or more *explicit justifications* as one's epistemic grounds for the belief in question. One might offer the justifications to other people as a public defense of one's belief, or one might "offer" them to oneself in private reasoning, or both. In any case, the offering of justifications must meet a minimum threshold of sincerity: A justification that involves a deliberate lie, for instance, is not a rationalization in our intended sense. One must in at least some weak sense, at least temporarily, accept that these justifications really do support one's belief, sufficiently so to make it rational to hold it.

In rationalization those cited justifications are sought only after the conclusion is already in hand. Rationalization is *post-hoc*. The conclusion is accepted or at least favored in advance, and one's desire to show the predetermined conclusion to be rational then motivates the search for explicit justifying grounds.

Post-hoc reasoning is often epistemically permissible. One believes something, or favors a conclusion, and then inquires as to its justificatory grounds. What is distinctive about rationalization in our intended sense is this: That post-hoc reasoning is guided by *a distorting factor*, something that leads one to favor the intended conclusion but which is not in fact good epistemic grounds for the conclusion, and which acts behind the scenes, as it were, to shape one's reasoning in an epistemically illegitimate way.

Thus, Dana's desire to keep the $20 leads her to favor the conclusion that it is permissible for her not to mention the seller's mistake. The same desire is also responsible for the excessively low and biased epistemic standards she uses in accepting a glib parade of justifications—that he's too foolish to deserve to be told, that that's just how the game is played, etc. Similarly, Mitchell's distaste for Kant leads him to favor the conclusion that Kant's arguments fail, without (let's suppose) being good epistemic grounds for thinking that Kant's arguments do in fact fail, and this dislike then operates strongly on his evaluation of justifications for rejecting Kant's arguments. Racial bias, an aversion to admitting wrongdoing, over-commitment to what you said yesterday, home team bias, etc., can all operate as distorting factors in a similar way.

In sum, rationalization is post-hoc reasoning toward a favored conclusion, where both the preference for the conclusion and the search for justifications are shaped by some epistemically non-probative distorting factor that isn't explicitly appealed to in those justifications.

We offer the following two counterfactual tests as diagnostic of rationalization:

> *Counterfactual Test A*: If the justifications offered had not been available to you, you would have sustained your approximate degree of confidence in P, either offering some other justification or abandoning the attempt to justify.

> *Counterfactual Test B*: If your preference had been absent, you would not have regarded those justifications as sufficient to render your approximate degree of confidence in P epistemically justified.[4]

On a natural interpretation of both Newspaper and Kant-Objector. Dana's and Mitchell's reasoning is better understood as an epistemically illegitimate attempt to justify a conclusion that is favored due to an epistemically non-probative factor than as an attempt to get at the truth whatever it might be. It is not primarily because they have evaluated their merit in an epistemically responsible way that Dana and Mitchell accept the justifications they offer; rather it's because they are so eager to establish the rationality of their favored belief. Had one justification not been available, they would have searched for others (Counterfactual Test A); and had they not been biased, they would not have been satisfied with the justifications they offered (Counterfactual Test B).

Tests A and B are neither necessary nor sufficient for rationalization in our intended sense. For instance, even highly motivated reasoners might be forced to abandon their preferred conclusions if they cannot find justifications that reach a minimum threshold of plausibility (Kunda, 1990). And even an inveterate rationalizer will sometimes land upon a powerful set of justificatory reasons, such that, even if she had evaluated those reasons in a sustained, epistemically responsible manner, she would have regarded them as sufficient to render her approximate degree of confidence in P epistemically justified. We intend Tests A and B as diagnostic of rationalization, rather than as criterial.

Rationalization admits of degrees and gray cases, along at least two dimensions. One concerns the degree of transgression. Suppose that without any particular preference for P, Miguel would be of the *slight* opinion that it is rational for him to believe P on the basis of reason R. However, Miguel does favor P, due to a distorting factor which biases his assessment of justifications, and consequently believe with *moderate* confidence that it is rational for him to believe P on the basis of reason R. His belief is a bit stronger than it ought to be. If Miguel is only a *little* more confident than he should be, then it seems not quite right to say that his reasoning has been "guided by" or "shaped by" the distorting factor sufficiently to count as full-on rationalization. We see no sharp line between cases like this and clear-cut cases like Newspaper.

The other type of gray case is temporal. In the cleanest cases of rationalization, the thinker has never before considered the issue at hand. Dana has never thought about whether it's wrong to keep the $20 in the sort of situation she is in. In other cases, the conclusion or some near relative of it, will have occurred to the thinker before, along with considerations pro and con. The thinker may even have accepted the conclusion before, perhaps for reasons other than the justification she now offers. In these cases, what determines whether the thinker's current reasoning involves rationalization is the extent to which her *current* preference for the conclusion is the result of a factor that does not constitute good epistemic grounds for it and subsequently taints the search for justifications.

Even in paradigmatic cases of rationalization, but especially in gray cases like these, it will often be difficult to ascertain to what extent the person in question did in fact rationalize (whether the person is you or someone else).

Sometimes it's epistemically fine to be "biased" toward a favored conclusion, if you favor the conclusion on good epistemic grounds and your bias is warranted. This is not rationalization in our intended sense. For example, if you were to read about a study by the Ice Cream Manufacturers' Advocacy Group showing that consuming large amounts of ice cream improves life expectancy by three years, you would presumably be justified in reading the study with a skeptical eye, looking closely for the flaws you anticipate it must have, and inferring the existence of a particular one on relatively light evidence. Although you favor a certain conclusion in advance ("the study is poor quality") and you are searching to justify that conclusion ("aha, probably healthier people were assigned to the ice-cream-eating group!"), it is not rationalization in our sense if there is no epistemically non-probative distorting factor at work behind the scenes.

Rationalization in moral reasoning

Psychological research suggests that rationalization is common in the moral domain. Consider Jonathan Haidt's famous "dumbfounding" studies. Participants are told the story of Mark and Julie, a brother and sister travelling together in Europe who decide to have sex once, just for fun, then never do so again. Mark and Julie use two forms of birth control, enjoy the experience, never tell anyone, and it strengthens their relationship with each other. Most participants judge that it was wrong for Mark and Julie to have had sex. As portrayed in Haidt (2012), participants reach for one justification after another. For example, a participant might start by mentioning the possibility of birth defects. When she is reminded that Mark and Julie used two forms of birth control, she might shift to saying it will harm their relationship. When this too is shot down, she tries something else until eventually she says she can't explain why it's wrong; she just knows it is wrong. Haidt argues that participants' attitudes are driven by an emotional or intuitive commitment to a norm of "purity" grounded in a sense of disgust. The results are controversial (Kennett, 2012; Railton, 2014; Royzman, Kim, & Leeman, 2015), but such cases invite interpretation as rationalization: That one finds an act sexually disgusting is poor epistemic grounds for thinking the act immoral, and thus a distorting factor (in our sense) on participants' reasoning; and this distorting factor then guides their post-hoc search for and evaluation of justifications toward the preferred conclusion. If it is the case that rationalization occurs in cases like this, it seems plausible that it would also be present in the more common cases where a post-hoc rationalizing process is able to find minimally adequate justifications after a brief and biased search.

The literature on implicit bias also suggests that rationalizations might be common.[5] Hodson, Dovidio, and Gaertner (2002) asked participants to evaluate samples of putative college applicants. Some applications had high college board scores and mediocre high school achievement (along with other information). Other applications had mediocre college board scores and excellent high school achievement. Some applicants were racially Black (based on photo) and others White. Participants who had previously scored high on a measure of aversive racism, most of whom would presumably disavow personal prejudice, tended to assess applications more negatively when the applicants were Black than when they were White. Participants were then asked to rank the relative importance, in college admissions, of high school achievement, college board scores, and other factors. High prejudice (but not low prejudice) participants tended to rate college board scores more important than high school achievement when the Black applicant had low board scores and excellent high school achievement, and they tended to rate high school achievement more important than board scores when the Black applicant had mediocre high school achievement and high college board scores. This pattern of results suggests that a substantial proportion of high-implicit-bias respondents may have rationalized in such cases—conveniently finding and endorsing justifications post-hoc out of a desire to justify their low assessment of the Black candidate.

In general, the literature on "dual process" theory suggests that many of our judgments arise from fast, intuitive "System 1" processes to which we have little or no introspective access, and which might reflect many kinds of epistemically undesirable bias.[6] A thinker's motives can have a significant impact on what memories, ideas, or facts even make it into consciousness for explicit consideration in the first place; and also on what argumentative strategies she pursues, the intensity of particular intuitions or feelings of confidence, and much more. The literature on cognitive dissonance suggests that we tend to accept and defend conclusions that make our previous choices look reasonable (Cooper, 2007; Festinger, 1957). The literature on positive self-illusions suggests that we are prone to over-estimate the nature and extent of our positive traits and the likelihood of positive outcomes in our future (Taylor & Brown, 1988). The literature on "myside" bias and "belief-overkill" suggests that on controversial issues we tend to recall and to generate more reasons on the side we favor than on the side we oppose, and that when making decisions we try to bring our beliefs into line so that not one of them counts against the option we prefer (Baron, 1995; Baron, 2009; Stanovich, West, & Toplak, 2013). We appear often to recruit explicit "System 2" processes after the fact to defend these biased intuitive judgments. If the System 1 sorts of factors that really drive our judgments are sufficiently epistemically non-probative and our recruitment of explicit reasoning in support of such judgments sufficiently biased by those factors, then such defenses are rationalizations.

All of this might happen outside of your awareness. You might have no idea how biased your reasoning is. You might think you are being entirely objective, conscientiously weighing up factors both pro and con in a perfectly even-handed way. The types of mechanisms here are generally thought to operate outside of conscious awareness. You simply reason as best you can, according to how things strike you, unaware of the biased mechanisms underneath.

Philosophers, ethicists, and cognitive scientists

We suspect that rationalization is common in the thought of philosophers and scientists, including on the topics of their expertise where the topics of their expertise are moral or philosophical. Our argument is this: (a) The topics of morality and philosophy are at least as ripe for rationalization as are most other topics, perhaps riper; so we should expect that people would commonly rationalize in thinking about these topics. (b) There is little reason to think that professional experts on these topics would do better: Empirical evidence suggests that neither high academic intelligence nor philosophical expertise is protective against rationalization; and in fact they might enhance the tendency to rationalize. (c) Nor does being reflective, introspective, and vigilant—as one might be upon learning more about the phenomenon of rationalization—appear to be effective in reducing rationalization. (d) Finally, anecdotal and historical evidence provide informal support for the idea.

People's moral and philosophical reasoning is ripe for rationalization

There is no straightforward way to measure how common rationalization is in everyday life—that is, no straightforward way to measure the frequency and proportion of explicit justification or reason-giving that is best explained as highly biased post-hoc justification-seeking of conclusions that are favored for epistemically non-probative reasons. Everyday experience and laboratory experiments suggest that it might be common; but scholars might reasonably disagree about how widespread rationalization is in everyday experience and about the extent to which the relevant laboratory studies apply to everyday life.

Two broad types of consideration support the idea that rationalization might be common especially in the moral and philosophical reasoning of ordinary people. One is just ordinary observation. The sorts of arguments one hears from distant relatives at a holiday dinner and that one sees in social media feeds often invite explanation in terms of rationalization. Why does your uncle reason in this way rather than that way about gun control, or climate change, or tax rates? What lies behind your colleague's legion of excuses not to serve on the necessary but time-consuming department committee? When you disagree, others' rationalizing patterns can be—or at least *seem*—evident. But of course rationalization might be just as common, but more difficult to see, in people whose conclusions you accept. (And also, of course, rationalization might not be present everywhere you think it is. Indeed, leaping quickly to the assumption that other people are rationalizing can itself be a kind of rationalizing justification for dismissing their views.)

Our second consideration is that moral and philosophical disagreements are especially fertile grounds for rationalization: People often care intensely about moral and philosophical issues; the positions rarely admit of straightforward external correction; and the issues offer many loci for bias to enter into the reasoning process.

Big-picture moral and philosophical questions are often very important even to ordinary non-philosophers. Is there an immaterial soul? Is it wrong to cheat on one's taxes? What obligations do we have or not have to impoverished people in distant lands (or in our midst)? People care about these questions. They are invested in certain answers, for emotional reasons, for reasons of cultural identity or personal self-conception, for self-serving reasons—or even just because they are attached to their presuppositions and after having given their first answer they like to stick with it. People are then strongly motivated to defend one side of the question.

Moral and philosophical questions typically admit of no straightforward proof or refutation, instead opening up into a complexity of considerations, which provide many opportunities for these preferences to have an impact. For instance, in philosophy many lines of thinking turn crucially on one's "sense of plausibility," as emphasized by Kornblith:

> [M]any arguments involve subtle appeals to plausibility. There can be little doubt that the rationalizer's sense of plausibility is affected in important ways by the motivation he has for rationalizing, and this does not aid in the project of coming to believe truths.
>
> *(1999, p. 185)*

Similarly, in ordinary everyday moral and philosophical reasoning, people can reach very different assessments of the *force* of a particular reason or the *significance* of a consideration. Reasons strike us as compelling, or not; similarities strike us as relevant or not. Objections to a position seem threatening, or to be mere cavils that could presumably be dealt with. The disagreement of peers can give us pause, leave us indifferent, or inspire a contrarian impulse to push farther. People's preferences might affect perceptual judgments (including size, color, and distance; Balcetis and Dunning, 2006, 2007, 2010); what is remembered (Kunda, 1990; Mele, 2001); what facts are called to mind, with what vividness and salience (Mele, 2001); what hypotheses are envisioned (Trope & Liberman, 1996); what one attends to, and for how long (Lord et al., 1979; Nickerson, 1998); and so forth—just about every phase and aspect of cognition. All of these assessments are potential loci for bias to enter. At each of them, we might unwittingly thumb the scales a bit toward our favored conclusion.

Methodological judgments also offer a range of loci for bias: to what extent do you trust scientific results, and which ones? Who should be considered an authority whose opinion deserves weight? What should you spend your time thinking about, and what isn't worth much consideration? What argumentative tacks do you explore? How much critical attention should you pay to your own beliefs, and their sources, and which ones, in which respects? How much trust should you have in your intuitive first judgments vs. more explicitly reasoned responses? How much trust should you invest in your feelings of confidence? Often these questions are answered only implicitly.[7]

Patterns of bias can compound across several questions, so that with many loci for bias to enter, the person who is only slightly biased (e.g., slightly more confident

in her belief than is warranted) on each of many questions can ultimately come to a very different position than would someone who was not biased in the same way. Rationalization could operate either by producing many small biases that cumulatively tilt you toward a conclusion you would not otherwise have reached or by influencing you mightily at a crucial step.[8]

There is little reason to think professional experts would be better

One might allow that ordinary non-philosophers commonly rationalize in considering moral and philosophical questions, recruiting justifications post-hoc in favor of conclusions antecedently favored due to a distorting factor of little justificatory epistemic relevance. But maybe professional experts in moral and philosophical reasoning would rationalize less? Professional philosophers and cognitive scientists are presumably more academically intelligent than the general population, and philosophers in particular might be unusually good at verbal reasoning (Kuhn, 1991). Perhaps these are protective against rationalization? Furthermore, people who reason regularly about moral and philosophical matters in a professional context might have specific disciplinary expertise on those topics that reduces the likelihood of rationalization.

Existing evidence on these questions is limited. But what evidence there is suggests that neither academic intelligence nor disciplinary expertise in philosophy or ethics is protective against rationalization. They may in fact enhance it.

Although academic intelligence and experience in verbal argumentation might enhance reasoners' ability to spot weak arguments, any such advantage might be counterbalanced or more than counterbalanced by an increased ability to discover arguments toward a favored conclusion. Stanovich, West, and Toplak (2013), reviewing several studies, found that the degree of myside bias is largely independent of intelligence or other measures of cognitive ability. Dan Kahan has found that on several measures people who use more "System 2" type explicit reasoning show higher rates of motivated cognition rather than lower rates (Kahan, 2011, 2013; Kahan et al., 2011). Furthermore, thinkers who are more knowledgeable will have more facts to choose from when constructing a line of motivated reasoning (Braman, 2009; Taber & Lodge, 2006). If a professional ethicist wants to steal a library book, for example (Schwitzgebel, 2009), she can no doubt discover some at least superficially plausible justification in terms of half a dozen different ethical theories, and she might be especially interested in doing so. Compared to more informal philosophical and moral reasoning, extended professional philosophical discussion plausibly offers at least as many loci for bias to enter—plausibility judgments, subsidiary moves, methodological presuppositions, historical appeals, etc. While some biases and biased processes are less prevalent among those who score high on standard measures of intelligence, others have been shown to be no less frequent or powerful; rationalization may be one of them.[9]

Nor does disciplinary expertise specifically in philosophy appear to be protective, at least based on a few studies from the lab of one of the authors of this

chapter. Schwitzgebel and Rust (2014), for instance, hypothesized that professional philosophers, and especially professional ethicists, would tend to show higher correlations between their expressed attitudes about moral issues and their self-reported or directly measured behavior on those same issues if either of the following two views is correct: a "booster" view on which ethicists discover moral truths and then shape their behavior to match those moral discoveries, or a "rationalization" view on which ethicists are especially prone to use their professional skills to construct or defend moral attitudes that match their pre-existing behavioral inclinations. Both hypotheses were disconfirmed: across a range of measures, Schwitzgebel and Rust found that ethicist philosophers, philosophers not specializing in ethics, and a comparison group of other professors from the same universities had similar low-to-moderate correlations between expressed moral opinions and self-reported or directly measured moral behavior. All groups appeared to rationalize at about the same rate, despite differences in topical academic expertise. Although this does not support the idea that philosophers (or other people who are more knowledgeable) rationalize more, it does speak against the idea that they rationalize less.

Schwitzgebel and Cushman (2012, 2015) presented moral dilemma scenarios to professional philosophers and two comparison groups of non-philosophers, followed by the opportunity to endorse or reject various moral principles. Professional philosophers were just as prone to irrational order effects and framing effects in their judgments about the scenarios as were the other groups, and were also at least as likely to "rationalize" their manipulated scenario judgments by appealing to principles post-hoc in a way that would render those manipulated judgments rational. Joshua Greene (2014) deploys these results as part of a general argument that philosophers with broadly deontological moral judgments (such as that you shouldn't push one person in front of a trolley, killing her to save five others) tend to deploy philosophical reasoning in an epistemically illegitimate rationalizing manner to justify their intuitive, emotional assessments.[10]

Reflection, introspection, and vigilance are not particularly protective against rationalization

Since the mechanisms of rationalization are largely non-conscious, one might not expect introspection to reveal them. And indeed the general finding in the psychological literature appears to be that people have a "bias blind spot": People tend to regard themselves as much less biased than other people, for example in their degree of self-serving bias and racial bias—sometimes even exhibiting more bias by objective measures the less biased they believe themselves to be (Pronin, Gilovich, & Ross, 2004; Uhlmann & Cohen, 2005). Indeed, efforts to reduce bias and be vigilant in spotting it could potentially amplify bias as follows: One examines one's reasoning for patterns of bias, finds no evidence of bias because of one's bias blind spot, and then inflates one's confidence that one's judgment or reasoning on that occasion is not biased: "I really *am* being completely objective and reasonable!" (as suggested in Ehrlinger, Gilovich, & Ross, 2005). People who have high estimates

of their objectivity might also be less likely to take corrective measures against bias (Scopelleti et al., 2015).

We don't reject the possibility that there are effective approaches to correcting for bias—perhaps especially approaches that involve increased exposure to counterarguments or alternative points of view, or making concrete predictions that can be falsified (Kahneman & Klein, 2009; Tetlock & Gardner, 2015). However, we doubt one should trust one's subjective assessment of the extent to which one is biased and prone to rationalization. Simply being reflective, introspective, and vigilant, in one's own judgment, is insufficient to warrant confidence that one is not rationalizing.

Observations from the history of philosophy

Nietzsche saw almost the entire history of philosophy as a history of rationalization:

> What provokes one to look at all philosophers half suspiciously, half mockingly . . . [is] that they are not honest enough in their work, although they make a lot of virtuous noise when the problem of truthfulness is touched even remotely. They all pose as if they had discovered and reached their real opinions through the self-development of a cold, pure, divinely unconcerned dialectic . . . while at bottom it is an assumption, a hunch, indeed a kind of "inspiration"—most often a desire of the heart that has been filtered and made abstract—that they defend with reasons they have sought after the fact. They are all advocates who resent that name.
>
> *(1886/1966, sec. 5, p. 12)*

Thus, according to Nietzsche, Spinoza's tuberculosis led him to emphasize self-preservation (1882/1974, sec. 349, pp. 291–292); the Stoics hypnotized themselves into seeing all of nature as reflecting their own image (1886/1966, sec. 9, pp. 15–16); the weak concoct justifications for thinking of their weakness as freely chosen moral virtue (1887/1998); etc.

Tolstoy writes:

> I know that most human beings—not only those considered clever but even those who are very clever and capable of understanding the most difficult scientific, mathematical, or philosophic problems—can very seldom discern even the simplest and most obvious truth if it be such as to oblige them to admit the falsity of conclusions they have formed, perhaps with much difficulty—conclusions of which they are proud, which they have taught to others, and on which they have built their lives.
>
> *(1896/1996, p. 131)*

It is untenable, we think, for a philosopher or scientist to maintain with confidence that his or her moral or philosophical reasoning is not substantially impacted by rationalization.[11]

Immanuel Kant is an interesting test case. He is one of the most respected moral philosophers in the history of philosophy. In his *Metaphysics of Morals* (1797/1991) he argues that masturbation is immoral in such a "high degree" that "in terms of its form" it "exceed[s] even murdering oneself" (pp. 221/425). He defends this claim by saying that in masturbating "a man surrenders his personality (throwing it away), since he uses himself as a means to satisfy an animal impulse" (p. 221/425). He argues that women and servants should not be allowed the right to vote since "their existence is, as it were, only inherence" (pp. 125–126/314–315). Bastard children can be freely killed since they "are born outside of the law (for the law is marriage) and therefore outside the protection of the law" (pp. 144–145/336). It's of course a matter of speculation what might have led Kant to favor these conclusions, but it's a salient possibility that distorting factors played a major role—classism, sexism, prudishness, etc.—and that Kant's arguments were substantially suited post-hoc to fit.

It is a matter of difficult judgment how Nietzschean to be in one's reading of the history of philosophy, how common the pattern is toward which Tolstoy points, or what lessons to draw from the case of Kant. We invite you to consider the possibility that the ethical and philosophical reasoning of even the very best philosophers is rife with rationalization. And then we invite you to turn your eye fearfully upon yourself.

Moral and philosophical thinking, in addition to sometimes *instantiating* rationalization, can—even if reasonable in a way at the moment—depend on *previous* rationalizations that remain unchallenged. For instance, today a philosopher infers Q from P. Her belief that P was formed long ago on the basis of rationalization. But her inference to Q on the basis of P, let us suppose, is not itself best explained by a desire to establish that Q is true. Even if she does desire to establish Q, what explains her inference is her sober recognition that P justifies Q. There is no mismatch between the justificatory grounds she offers for her belief that Q and the underlying causes of that belief. All the same, her belief that Q is the result of her past and uncorrected rationalization of P. In Newspaper, Dana rationalizes her way to the conclusion that it was okay for her to keep the change. Remembering this when she later sees Julio keep extra change given by a café worker, Dana might, in a non-rationalizing way, reason that since it was okay for her to keep the newspaper change, and there's no relevant difference between the situations, it's also okay for Julio to keep his change. Dana would not have reached this conclusion about Julio had she not earlier rationalized in her own case. Dana's inference about Julio is not itself a rationalization, but rather involves a belief that originated in a past rationalization.

Some considerations against the pervasiveness of rationalization in moral and philosophical reasoning

We think it unlikely that *all* moral and philosophical reasoning either is or depends upon rationalization in our sense of the term. Questions that admit of straightforward

formal solutions, for example, and technical sub-questions on which one has no antecedently preferred opinion, seem to offer fewer loci for rationalization and less motive for it. Also, it appears that sometimes people are convinced by philosophical and moral arguments very much against their initial inclinations and desires. One possible example of this is people who have been convinced by Peter Singer's (1972, 2009) arguments that members of the upper and middle classes should donate most of their wealth to charity.[12] Few of the philosophers who have been convinced by Singer's arguments, we suspect, *wanted* antecedently to be convinced. Instead, most presumably would have preferred to find it morally permissible to continue enjoying their iPhones and restaurant meals.

Another important class of non-rationalizations are what we might call *basis shifting* cases. In basis shifting, one starts out highly biased, embracing a conclusion without good epistemic warrant, then one explores the issue and fortuitously finds fully satisfactory grounds for one's initial biased opinion—grounds which then become the new basis of one's opinion. Philosophers who recognize bias and rationalization in their past might defend the rationality of their current views, even if largely unchanged, by claiming that they have shifted basis. In some cases, this might even be reasonable.

So what?

Suppose that lots of moral and philosophical thinking does involve or depend upon rationalization. An interesting epistemic question is, so what? Is rationalization in fact epistemically bad?

Here are two reasons one might think rationalization may not be so bad:

(1) Moral and philosophical reasoning is a group enterprise, and *the community benefits from having passionate advocates with a wide variety of opinions, who defend and pursue their ideas come what may*. Even if some of those people fail to be epistemically rational at the individual level, they might contribute to epistemic rationality at the group level. Maybe the scientific psychological community, for example, needs people who passionately support currently unpopular versions of nativism and empiricism, even against the best overall weighting of evidence, giving those views the best defense they can muster. Moral and philosophical communities might likewise benefit from passionate advocates of unlikely or disvalued positions. (See Kuhn, 1962/1970, and Longino, 1990, on scientific communities.)

(2) *Even if rationalization is not epistemically beneficial, it might not be deleterious, at least in the context of professional philosophy*. Who cares why a philosopher has the views she does? All that matters, one might think, is the quality of the arguments that are produced. Low-quality arguments will be quickly shot down, and high-quality arguments will survive even if their origins are epistemically problematic. To use a famous scientific example: It doesn't matter if a vision of the structure of benzene came to you in a dream, as long as you can defend your view of that structure after the fact, in dialogue with your peers. (See Popper, 1934/1959.)

There can also be prudential, hedonic, and interpersonal advantages to rationalization. Rationalization might increase happiness or well-being (Du Châtelet c. 1740/1997; Erasmus, 1509/1989; Taylor & Brown, 1988). It might also help us strategically in influencing others (Mercier & Sperber, 2011; Trivers, 2000, 2010).

While acknowledging these points, we think that the epistemic costs of rationalization outweigh the epistemic benefits:

(A) *Rationalization leads to overconfidence.* If one favors conclusion P for reasons that aren't in fact good grounds for believing P, and then systematically pursues and evaluates evidence concerning P in a highly biased manner, one will often settle upon an unwarranted belief. One might conclude that P despite the preponderance of the available evidence supporting the opposite of P. Alternately, even when the evidence warrants a belief in P, one might end up believing P with more *confidence* than is warranted. This lesser transgression can be quite dangerous when one is, say, deciding whether to convict the defendant, upbraid the student, or do a morally questionable action. It can also substantially influence the subsequent assimilation of new information. (See Ellis [manuscript-a] for a more detailed discussion of the nature, frequency, and epistemic significance of these sorts of lesser transgressions.)

(B) *Rationalization impedes peer critique.* There's a type of dialectical critique that is, we think, epistemically important in moral and philosophical reasoning—we might call it *engaged* or *open* dialogue—in which one aims to offer to an interlocutor, for the interlocutor's examination and criticism, one's *real reasons* for believing some conclusion. One says, "here's why I think P," with the aim of offering considerations in favor of P that simultaneously play two roles: (i) they epistemically support P (at least prima facie); and (ii) acceptance of them is actually causally effective in sustaining one's belief that P is the case. Exposing not only your conclusion but your reasons for favoring that conclusion offers your interlocutor two entry points for critique rather than just one: not only "is P true or well supported?" but also "is *your* belief that P is well supported by the grounds you appeal to?" These can come apart, especially in the case where your interlocutor might be neutral about P but rightly confident that your basis for belief is insufficient. ("I don't know whether the stock market will rise tomorrow, but seeing some guy on TV say it will rise isn't good grounds for believing it will.") Rationalization disrupts this type of peer critique. One's real basis remains hidden; it's not really up for peer examination, not really exposed to the risk of refutation or repudiation. If one's putative basis is undermined one is likely simply to hunt around for a new putatively justifying reason. (Compare Habermas on sincerity and truthfulness in discourse ethics: 1981/1984, 2003.)

(C) In an analogous way, *rationalization undermines self-critique.* An important type of self-critique resembles peer critique. One steps back to explicitly consider

one's putative reasons for believing P, with the idea that reflection might reveal them to be less compelling than one had initially thought. As in the peer case, if one is rationalizing, the putative reasons don't really explain why one believes, and one's belief is likely to survive any potential undercutting of those putative reasons. The real psychological explanation of why you believe remains hidden, unexamined, not exposed to self-evaluation. (See Burge, 2013; Shoemaker, 1988; Williams, 2004. We say this despite agreeing with Carruthers, 2011, Cassam, 2014, and Kornblith, 2012, that people might not attain this type of self-knowledge nearly as often as they suppose.)

(D) *Rationalization engenders distrust and testimonial injustice.* Testimony, broadly construed, is integral to intellectual progress, but its value can be acutely compromised as a result of the "bias blind spot" we mentioned earlier. In discussions, people will tend to see more rationalization in others than in themselves. This can have significant epistemic consequences, especially when the disagreement concerns a belief or value central to a person's identity. For instance, Sara might regard Camila's rationalization (which Sara sees but Camila doesn't) as reason to be wary of Camila's credibility and epistemic practices—a wariness that might increase when Sara learns that Camila thinks *Sara* is the one who is rationalizing. That increase might then be visible to Camila, increasing *Camila's* distrust in turn. And so on, in a self-reinforcing cycle (Kahan, 2011). What results is an inaccurate conception on both sides of the relative value and credibility of the other as an epistemic agent and resource, a distortion with both epistemic and social ramifications.[13]

It is also unclear how much comfort is really justified by consideration (2), concerning quality detection. In moral and philosophical reasoning, quality can be difficult to assess. We are not confident that a philosophical community full of rationalizers is likely to reject only low-quality arguments, especially if patterns of motivated reasoning don't scatter randomly through the community, but tend to favor certain conclusions over others for reasons other than epistemic merit.

To be sure, it can be valuable to engage in post-hoc reasoning to substantiate views one finds intuitively plausible for unknown reasons. If one is justified in being highly confident in the reliability of one's intuition about something (that it's wrong to kill an unwilling person to harvest her organs to save five others, for instance), one may be justified in then hunting around post-hoc for an adequate justifying argument. But this is not rationalization. To be rationalization in our sense of the term, post-hoc reasoning must be guided by a distorting factor: a factor that causes (but does not justify) one's initial preference for the conclusion and subsequently also guides the search for justifications in an epistemically illegimate way. When a belief in the reliability of one's intuition about something is epistemically justified, it is not a distorting factor. A commitment to the value of intuition is not a commitment to the value of rationalization. Of course, ascertaining the epistemic grounds of one's trust in one's intuition can be quite difficult,[14] and a person's confidence in the reliability of her intuition can be *unjustified*, perhaps the result of a past rationalization, or a present rationalization, or something else entirely. Similarly, a confident belief in the reliability of one's intuition about something can be justified yet false.[15]

For these reasons we think we ought to be disappointed and concerned if it turns out that our moral and philosophical reasoning to a large extent either is or crucially depends upon post-hoc rationalization.

What can we do?

Suppose that rationalization is both epistemically undesirable and widespread in moral and philosophical thought. Is there anything we can do about it? Two issues are forward-looking: What, if anything, can we do to reduce the extent to which we rationalize? And what, if anything, can we do to mitigate the effects of rationalization when it happens? A third issue concerns the present: To what extent, if at all, should we reduce the confidence we have today in our moral and philosophical beliefs, and which ones?

The last of these three issues connects closely with recent debates in epistemology about peer disagreement. The peer disagreement literature asks: When and to what extent is a thinker being irresponsibly dogmatic in maintaining confidence upon learning that an "epistemic peer" disagrees with her? The epistemology of rationalization can be framed in a similar way: When and to what extent is a thinker being irresponsibly dogmatic in not budging as she begins to appreciate the potential reach of rationalization?[16] As for the other two issues, some research suggests that exercises of "self-affirmation" can reduce or pre-empt defensiveness and rationalization (Cohen et al., 2007; Critcher, Dunning, & Armor, 2010), although it's an open question how well that would transfer to the practice of professional philosophy and moral psychology. Another corrective worth exploring is for researchers to be more explicit about their vested interests, their preferences, their motivations—about what they would prefer not to have to conclude. Writing in dialogue with a respected opponent might also reduce rationalization, both in the authors (whose rationalizations might be exposed and who would be forced to avoid quick, uncharitable dismissals of their opponents' views) and maybe even in readers who would see this respectful engagement.

At the community level, rationalization is especially harmful when participants in the conversation have similar initial starting points or views they find attractive. While engagement with competing perspectives can sometimes *increase* polarization, and precisely because of the phenomena we have been discussing, we think that strengthening intellectual diversity in community discussion might nonetheless help limit the negative epistemic effects of rationalization. A better understanding of these phenomena could potentially have major social as well as epistemic benefits. The importance of female, minority, and other disvalued voices in academic dialogue is of course a recurrent theme in gender studies, ethnic studies, queer studies, and disability studies (e.g., Fricker, 2007; Longino, 1990; Medina, 2013). If rationalization is common in philosophy, that might afford yet one more reason to encourage efforts to substantially broaden the range of voices that are heard.[17]

Notes

1 The order of authorship was determined arbitrarily; both authors contributed equally to the manuscript. The arguments in the fourth section, concerning the potential reach of rationalization, stem largely from Jon's arguments in Ellis (manuscript-a) concerning the broader family of motivated reasoning.

2 The names in our examples were chosen after the examples were written, drawn randomly from lists of first names of former students in large lower-division classes at our universities. We hope that randomized name selection procedures will in the long run reduce bias and improve cultural representativeness. To avoid confusion or offense, we excluded "Jesus," "Mohammed," and very uncommon names.

3 We have a standard, pejorative sense of "rationalization" in mind. For two altogether different senses of the term, see Davidson, 1963 and (translations of) Weber, 1904/1905.

4 Counterfactual Test B resembles the "impartial observer" test for self-deception in Mele, 2001, pp. 106–107.

5 See Seigel, 2014 for a similar discussion of the epistemology of rationalization in implicit bias, based on data from Ulhmann and Cohen, 2005. For a general view of the literature on implicit bias, see Brownstein, 2015.

6 On dual-process models of cognition see Evans, 2007; Evans and Frankish, 2009; Kahneman, 2011.

7 We recognize that our reasoning in this chapter contains many of these same loci for potential rationalization. For instance, although each of us finds experimental results in social and cognitive psychology to be of significant import for the issues we are discussing, judgments about the force and significance of such studies are precisely the type of judgments in which rationalization might play a substantial role.

8 For more on the widely varied epistemic junctures at which a motive can have a surreptitious effect on philosophical reasoning, and their potential collective impact upon sincere, reflective, and intelligent thinkers, see Ellis (manuscript-a).

9 For a review of empirical work on the relation between particular biases and various measures of intelligence and cognitive ability, see Stanovich, 2011.

10 Consequentialist reasoning, Greene argues, works in a manner less readily describable as rationalization. We take no stand on that question.

11 The impact of psychological forces on philosophical reasoning is a pervasive theme in Ludwig Wittgenstein's philosophy as well. Wittgenstein was perpetually grappling with their impact on his own thinking: "I always want to bargain down the truth that I know & when it is unpleasant & again and again have thoughts with which I want to deceive myself" (2003, p. 217). For discussion, see Ellis and Guevara, 2012.

12 We thank Henry Shevlin for the example.

13 For a discussion of the rich network of bi-directional relationships between epistemic "injustice" and social injustice, see Medina, 2013. Testimonial injustice, for instance, often results from epistemic resistances that "function as obstacles, as weights that slow us down or preclude us from following (or even having access to) certain paths or pursuing further certain questions, problems, curiosities" (Medina, 2013, p. 48).

14 And for a variety of reasons. For one thing, one's search for those grounds can itself be susceptible to rationalization, if indeed one is motivated to find them. For another, our reasons for our beliefs are often inaccessible to us in important ways. As Ernest Sosa writes: "We have reasons . . . that, acting in concert, across time, have motivated our present beliefs, but we are in no position to detail these reasons fully" (Sosa, 2010, p. 291).

15 These are among the possibilities that can complicate attributions of rationalization, such as to Kant's reasoning about masturbation or Haidt's subjects' reasoning about incest. See Kennett, 2012 and Railton, 2014.

16 See Ellis (manuscript-b) for a sustained investigation of the epistemic significance of motivated reasoning and its close connection to the debate about disagreement. See also Avnur and Scott-Kakures (2015). On peer disagreement, see Feldman and Warfield, 2010.

17 For helpful discussion we thank Maudemarie Clarke, Georgia Warnke, and the many people who commented on relevant posts on *The Splintered Mind* blog, Eric's Facebook page, and other social media sites.

References

Audi, R. (1985). Rationalization and rationality. *Synthese, 65*, 159–184.

Avnur, Y., & Scott-Kakures, D. (2015). How irrelevant influences bias belief. *Philosophical Perspectives, 29*, 7–39.

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology, 91*, 612–625.

Balcetis, E., & Dunning, D. (2007). Cognitive dissonance and the perception of natural environments. *Psychological Science, 18*, 917–921.

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science, 21*, 147–152.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning, 1*, 221–235.

Baron, J. (2009). Belief-overkill in political judgments. *Informal Logic, 29*, 368–378.

Braman, E. (2009). *Law, politics, and perception: How policy preferences influence legal reasoning.* Charlottesville, VA: University of Virginia Press.

Brownstein, M. (2015). Implicit bias. *Stanford Encyclopedia of Philosophy* (Spring 2015 edition). Retrieved from http://plato.stanford.edu/archives/spr2015/entries/implicit-bias.

Burge, T. (2013). *Cognition through understanding.* Oxford, UK: Oxford University Press.

Carruthers, P. (2011). *The opacity of mind.* Oxford, UK: Oxford University Press.

Cassam, Q. (2014). *Self-knowledge for humans.* Oxford, UK: Oxford University Press.

Cohen, G. L., Sherman, D. K., Bastardi, A., Hsu, L., McGoey, M., & Ross, L. (2007). Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *Journal of Personality and Social Psychology, 93*, 415–430.

Cooper, J. (2007). *Cognitive dissonance: 50 years of a classic theory.* London: Sage.

Critcher, C., Dunning, D., & Armor, D. (2010). When self-affirmations reduce defensiveness: Timing is key. *Personality and Social Psychology Bulletin, 36*, 947–959.

Davidson, D. (1963). Action, reasons and causes. *Journal of Philosophy, 60*, 685–700.

Du Châtelet, É. (1747/2008). *Discours sur le bonheur.* In R. Mauzi, *L'art de vivre d'une femme au XVIIIe siècle.* Paris: Desjonquères.

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: People's assessments of bias in themselves and others. *Personality and Social Psychology Bulletin, 31*, 680–692.

Ellis, J. (manuscript-a). The reverberating impacts of motivated reasoning-lite.

Ellis, J. (manuscript-b). Disagreement and motivated reasoning.

Ellis, J., & Guevara, D. (2012). Introduction. In J. Ellis & D. Guevara (Eds.), *Wittgenstein and the philosophy of mind.* New York: Oxford University Press.

Erasmus. (1509/1989). *The praise of folly and other writings* (R. M. Adams, Ed., Trans.). W.W. Norton.

Evans, J. S. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment.* Hove, UK: Psychology Press.

Evans, J. S. B. T., & Keith Frankish (2009). *In two minds.* Oxford, UK: Oxford University Press.

Feldman, R., & Warfield, T. A. (Eds.) (2010). *Disagreement.* New York: Oxford University Press.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford University Press.

Freud, S. (1911). Psycho-analytic notes on an autobiographical account of a case of paranoia (dementia paranoides). In *The standard edition of the complete psychological works of Sigmund Freud, volume XII (1911–1913): The case of Schreber, papers on technique and other works.*

Fricker, M. (2007). *Epistemic injustice*. Oxford, UK: Oxford University Press.
Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics*, *124*, 695–726.
Habermas, J. (1981/1984). *The theory of communicative action, vol. 1*. (T. McCarthy, Trans.). Boston, MA: Beacon.
Habermas, J. (2003). *Truth and justification*. Cambridge, MA: MIT.
Haidt, J. (2012). *The righteous mind*. New York: Pantheon.
Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin*, *28*, 460–471.
Jones, E. (1908). Rationalization in every-day life. *Journal of Abnormal Psychology*, *3*, 161–169.
Kahan, D. (2011). Neutral principles, motivated cognition, and some problems for constitutional law. *Harvard Law Review*, *125*, 1–77.
Kahan, D. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, *8*, 407–424.
Kahan, D. M., Wittlin, M., Peters, E., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. N. (2011). The tragedy of the risk-perception commons: Culture conflict, rationality conflict, and climate change. Temple University Legal Studies Research Paper No. 2011–26; Cultural Cognition Project Working Paper No. 89; Yale Law & Economics Research Paper No. 435; Yale Law School, Public Law Working Paper No. 230. Available at SSRN: http://ssrn.com/abstract=1871503 or http://dx.doi.org/10.2139/ssrn.1871503.
Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*, 515–526.
Kant, I. (1797/1991). *The metaphysics of morals* (M. Gregor, Trans.). Cambridge, UK: Cambridge University Press.
Kennett, J. (2012). Living with one's choices. Moral reasoning in vitro and in vivo. In R. Langdon & C. Mackenzie (Eds.), *Emotions, imagination, and moral reasoning*. New York: Taylor & Francis.
Kornblith, H. (1999). Distrusting reason. *Midwest Studies in Philosophy*, *23*, 181–196.
Kornblith, H. (2012). *On reflection*. Oxford, UK: Oxford University Press.
Kuhn, D. (1991). *The skills of argument*. Cambridge, UK: Cambridge University Press.
Kuhn, T. S. (1962/1970). *The structure of scientific revolutions*, 2nd ed. Chicago, IL: University of Chicago.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.
Longino, H. E. (1990). *Science as social knowledge*. Princeton, NJ: Princeton University Press.
Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.
Medina, J. (2013). *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and resistant imaginations*. Oxford, UK: Oxford University Press.
Mele, A. (2001). *Self-deception unmasked*. Princeton, NJ: Princeton University Press.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–111.
Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.
Nietzsche, F. (1882/1974). *The gay science* (W. Kaufmann, Trans.). New York: Random House.

Nietzsche, F. (1886/1966). *Beyond good and evil* (W. Kaufmann, Trans.). New York: Random House.
Nietzsche, F. (1887/1998). *On the genealogy of morality* (M. Clarke & A. J. Swensen, Trans.). Indianapolis, IN: Hackett.
Popper, K. (1934/1959). *The logic of scientific discovery*. London: Hutchinson.
Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*, 781–799.
Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics*, *124*, 813–859.
Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision Making*, *10*, 296–313.
Schwitzgebel, E. (2009). Do ethicists steal more books? *Philosophical Psychology*, *22*, 711–725.
Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, *27*, 135–153.
Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise, and reflection. *Cognition*, *141*, 127–137.
Schwitzgebel, E., & Rust, J. (2014). The moral behavior of ethics professors: Relationships among self-reported behavior, expressed normative attitude, and directly observed behavior. *Philosophical Psychology*, *27*, 293–327.
Scopelleti, I., Morewedge, C. K., McCormick, E., Min, H. L., Lebrecht, S., & Kassam, K. S. (2015). Bias blind spot: Structure, measurement, and consequences. *Management Science*, *61*, 2468–2486.
Shoemaker, S. (1988). On knowing one's own mind. *Philosophical Perspectives*, *2*, 183–209.
Siegel, S. (2014, November 7). Rationalization, belief, and inference [Blog post at *The Brains Blog*]. Retrieved from http://philosophyofbrains.com/2014/11/07/rationalization-belief-and-inference.aspx.
Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, *1*, 229–243.
Singer, P. (2009). *The life you can save*. New York: Random House.
Sosa, E. (2010). The epistemology of disagreement. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Social epistemology*. New York: Oxford University Press.
Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.
Stanovich, K. E., West, R. R., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*, 259–264.
Summers, J. S. (manuscript). Rationalization and its discontents.
Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*, 755–769.
Taylor, S. E., & Brown, J. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*, 193–210.
Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. New York: Crown.
Tolstoy, L. (1896/1996). *What is art?* Indianapolis, IN: Hackett.
Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences*, *97*, 114–131.
Trivers, R. (2010). Deceit and self-deception. In P. M. Kappeler & J. B. Silk (Eds.), *Mind the Gap*. Berlin: Springer-Verlag.
Trope, Y., & Liberman, A. (1996). Social hypothesis-testing: Cognitive and motivational mechanisms. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles*. New York: Guilford Press.

Uhlmann, E., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16*, 474–480.
Weber, M. (1904–05/1992). *The Protestant ethic and the spirit of capitalism* (T. Parsons, Trans.). London: Routledge.
Williams, M. (2004). Is knowledge a natural phenomenon? In R. Schantz (Ed.), *Volume 2 The externalist challenge*. Berlin: De Gruyter.
Wittgenstein, L. (2003). *Ludwig Wittgenstein: Public and private occasions*. In J. C. Klagge & A. Nordmann (Eds.). Lanham: Rowman & Littlefield.

11

EXILE OF THE ACCIDENTAL WITCH

Character and intention in an uncertain social world

Tage S. Rai

Abstract

Moral psychology is shifting from an act paradigm wherein actions are judged in isolation toward a character paradigm geared toward inferences of an actor's potential as a social-relational partner in the future. From this perspective, seemingly irrational influences of character become rational inputs into moral judgment, and intentions only matter when they have predictive validity for future social relations. Intentions are discounted when actors have pre-existing negative characteristics that preclude them from social relations, when negative actions cannot be prevented in the future regardless of intention, and when the damage done by actions cause irreparable damage to social relationships.

Introduction

Why is Mother Theresa more revered than Bill Gates, even though the Gates' foundation has saved considerably more lives (Pinker, 2008)? Why are people judged more severely if they have character flaws, even when those character flaws have no bearing on the crime that was committed (Nadler, 2012)? Why are people punished for events that were clearly unintended, as in honor killings of women after they have been raped (Zoepf, 2007), or in the exile of those accused of unconscious witchcraft (Evans-Prichard, 1937)?

Once upon a time, moral psychologists assumed that people make moral judgments by judging the rightness or wrongness of someone's actions and their mental states *in isolation*. In this isolated act paradigm, people are morally blameworthy if they have acted intentionally to cause harm to another person when they reasonably could have done otherwise in that particular instance. And indeed, experimental evidence has typically found that people are blamed or punished less severely if they acted unintentionally, even if the action itself remains identical (Cushman, 2008; Darley, Klosson, & Zanna, 1978; Gray, Young, & Waytz, 2012; Hauser, 2006; Mikhail, 2007).