FACTITIOUS VIRTUE

By

MARK ALFANO

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2011

This manuscript has been read and accepted for the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

	Michael Levin
_____	_____
Date	Chair of Examining Committee
	Iakovos Vasiliou
_____	_____
Date	Executive Officer

Jesse Prinz
_____

Rohit Parikh
_____

Michael Levin
_____
Supervision Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

FACTITIOUS VIRTUE

by

Mark Alfano

Advisor: Professor Jesse Prinz

The primary aim of this project is to argue that empirical challenges to moral theories like virtue ethics should be co-opted rather than resisted. Virtue ethics has much to offer. Its vision of a flourishing life seems a better object of moral contemplation and evaluation than the desiccated, ascetic rules of deontology and consequentialism; its focus on "thick" concepts like honesty and courage seem to bridge the *is/ought* gap; its weaving together of reasons and motivations obviates concerns about moral schizophrenia. Furthermore, the virtue ethical account of action paints a detailed picture of sensitivity to reasons, careful and correct construal of ambiguous information, and thoughtful deliberation.

Recently, however, philosophers informed by the situationist tradition in social and experimental psychology have begun to question the empirical presuppositions of virtue ethics, and a cottage industry has grown up around attacking and defending their arguments. To move the debate forward, I develop a comprehensive list of the empirical presuppositions of virtue ethics, the most contentious items of which are

consistency (if an agent possesses a virtue sensitive to reason *r*, then she responds to *r* whenever it is relevant), *explanatory power* (if an agent possesses a virtue, then reference to that virtue sometimes explains her behavior), *predictive power* (if an agent possesses a virtue, then reference to that virtue sometimes enables prediction of her behavior), and *egalitarianism* (almost anyone can be virtuous).

The empirical challenge relates to the conjunction of these four claims. Several decades of studies in social psychology have shown that most people sometimes respond not to the reasons there are for them to act on a given occasion but to situational factors like ambient sounds, ambient smells, moods and emotions, and presence of bystanders. These morally irrelevant but causally powerful factors can be unified under the heading of attentional focus: loud and annoying sounds, unpleasant smells, negative moods and emotions, and the presence of bystanders all lead to the focusing of attention on a small number of situational features, while their opposites lead to the dilation of attention. When their attention is focused, people exhibit inattentional blindness, which leads them to miss or misconstrue important moral features of their situations; insensible to the reasons there are for them to act, they deliberate poorly (if at all) and act in violation of virtue.

Situationist psychology does not just deny character traits. It also explains away the strongly felt intuition that there are character traits, invoking a virtual pantheon of gods of error and ignorance that includes the power of (mis)construal (misinterpreting ambiguous information as evidence for character traits), selection bias (using non-representative samples of behavior, thus overlook cases where people act in violation of

traits), availability bias (assuming that first impressions are representative), and confirmation bias (seeking and using only evidence that confirms first impressions). These mechanisms guarantee that intuition would lead us to believe in traits even if traits did not exist. Since one does not know that *p* if one would believe *p* were it false, we cannot know on the basis of intuitions that character traits exist.

Three primary responses to the situationist critique can be identified in the literature. *The dodge*: virtue is a rare ideal, so data showing that most people are not virtuous is moot. *The counterattack*: the data do not support the situationist critique. *The retreat*: although the situationist critique shows that global traits do not exist, a naturalistic theory of virtue can still be formulated in terms of actions or local traits. Unfortunately, most versions of these arguments are either unsound or give up the consistency, explanatory power, predictive power, or egalitarianism of virtue. The dodge, for instance, is an outright denial of egalitarianism. Most versions of the counterattack fail to individuate virtues by their characteristic reasons, and thus are morally inadequate; others rely on intuitions, which have already been shown unreliable. Two compatible tactics for dealing with the challenge, however, do emerge from this literature: an emphasis on the portability of context and a shift from situation-consumerism to situation-producerism. By recognizing the power of situations and identifying the types of situations (not) conducive to behavior in accordance with virtue, one can strategically seek (avoid) situations likely to lead to (non-)virtuous actions. And by recognizing the causal dialectic between agents and situations, one can shift to thinking of agents as active producers rather than mere passive consumers of situations

– a point of view that encourages the creation of situations conducive to behavior in accordance with virtue.

Along these lines, I argue that virtue (though not vice) attributions of the right sort should be made regardless of their truth-value. Drawing on formal work in multi-agent epistemic logic and empirical studies in social psychology, consumer research, and behavioral economics, I show that the plausible, public attribution of virtuous traits induces both identification with those traits and belief that others expect one to act in trait-consonant ways, which in turn leads to trait-consonant behavior. The notions of placebo effects and self-fulfilling prophecies are instructive parallels to virtue-labeling. Thinking of virtue attributions merely as fact or fiction is too limited. We must recognize in addition a third category: factitious attributions, which make themselves true by being plausibly, publicly announced.

In another example of the portability of context and situation-producerism, I present a novel theory of social distance in terms of *potential for interaction*, *group identity*, and *information*. This theory draws support from recent work by experimental psychologists and economists, as well as an experiment that I myself conducted. By manipulating heuristics that track social distance, agents can be led systematically to underestimate it, which in turn leads to elevated levels of behavior in accordance with virtue.

Acknowledgments

Nothing of value is accomplished alone. I would like to thank the many people who helped me in many ways during the planning, research, writing, and revising of this manuscript. The exigencies of memory may prevent me from thanking everyone who deserves it, but I hope at least to thank everyone I do mention as well as they deserve. First in these acknowledgments and first in my heart is my lovely wife Veronica, light of my life and fire of my loins. She somehow managed to keep both of us sane while we simultaneously wrote dissertations in the same 60 square-foot office. My parents, Ronald and Marjorie, my in-laws, Mary and David, and my dwarf silver marten, Nori, all provided moral support.

Many thanks as well to my advisor, Jesse Prinz, for his sage advice, pointed criticism, and unflagging encouragement. Without his responses to emails, guidance during office hours, and helpful questions at conferences and colloquia, many parts of this project would lack what rigor and vivacity they have attained. His influence is especially notable in sections 5, 7, and 8.

In addition to my advisor, I would like to express my warmest thanks to the other members of my committees – to Richard Sorabji and Catherine Wilson for their insightful criticisms of the master's thesis out of which this project grew; to Graham Priest and Samir Chopra, who graciously agreed to serve on my prospectus committee; to Rohit Parikh and Michael Levin, the two CUNY faculty who served on my prospectus and dissertation committees; and to John Doris and Gilbert Harman, who lit the fire that

illuminates much of my view and makes the empirical commitments of ethical theories such a burning question. Professors Levin and Parikh deserve my special thanks for their timely and incisive criticisms. Professor Parikh's influence is most evident in section 8, which refers directly to his work on knowledge-based obligation. Professor Levin's influence is more diffuse but no less important; thanks to him especially for challenging me on the empirical evidence against character traits and against preferential and hedonistic indeterminacy. Professors Doris and Harman also deserve exceptional thanks for their willingness to correspond with me about this project almost from its inception, and for agreeing to serve on the dissertation committee.

In addition to the members of my various committees, several others provided invaluable advice during various stages of this project. David Rosenthal helped me to construct the prospectus and dissertation committees, in addition to voicing his concerns over the reliability of evidence from social psychology. Joshua Knobe's skeptical questions on 15 March 2010 are the only begetter of section 8.3.1.

More distal but no less important influences on this work are Alexander Nehamas and David Murphy. Thanks to Professor Nehamas for advising my Princeton Senior Thesis on Nietzsche and Wittgenstein, for encouraging me to apply a second time to graduate schools when I crapped out the first time, and for cueing me into the secondary speech-acts that can be performed by virtue attributions. Thanks to Mr. Murphy for piquing my interest in philosophy when I was but a sophomoric high-schooler, and for staying in touch all these years afterwards. Nietzsche says that one

repays a teacher poorly if one remains only a student. I hope this dissertation prompts him to rejoin that one repays a student poorly if one remains only a teacher.

Many thanks as well to Brian Robinson, Daniel Shargel, Myrto Mylopoulos, and Todd Beattie for participating in the informal dissertation-completion group that helped us all to schedule our chapters, mark our progress, and test our ideas.

Thanks to my dear friends Jose Montes, Kim Thornton, Katie DiSalvo, Sarah Fine, and Julia Mansfield, as well as my siblings Edward, Shelley, and Bean. Thanks to Mark Greenstein; tutoring for Ivy Bound pretty much financed my entire graduate school career.

Finally, thanks to John Cleese, Jon Stewart, and the cast and writers of *Battlestar Galactica*. I couldn't have done it without you.

Table of Contents

List of tables

List of figures

Philosophers have hitherto only interpreted the world in various ways; the point is to change it.

~ Karl Marx, *Theses on Feuerbach*

Virtue is not left to stand alone. He who practices it will have neighbors.

~ Confucius, *Analects*

Chapter 1. Introduction

How do I know you're one of the good guys?
You dont. You'll have to take a shot.
Are you carrying the fire?
Am I what?
Carrying the fire.
You're kinda weirded out, arent you?
No.
Just a little.
Yeah.
That's okay.
So are you?
What, carrying the fire?
Yes.
Yeah, we are.
~ Cormac McCarthy, *The Road*

Gotham is a city rife with tales of virtue and vice.

On March 13, 1964, Kitty Genovese was raped and stabbed in the Kew Gardens neighborhood of Queens. Although dozens of neighbors heard Kitty's cries for help, no one called the police until at least half an hour after the assault began. In addition to decrying her murderer as a vicious killer, the public responded to the event with outrage at the apathetic witnesses. Did not one of them have an ounce of compassion?

On January 2, 2007, Cameron Hollopeter suffered a seizure and stumbled off the 137th-Street subway platform into the path of an oncoming train. One bystander, Wesley Autrey, noticed the emergency and dove onto the tracks. Lacking time to lift the victim back onto the platform, he pinned Hollopeter in the drainage trench between the rails while the train straddled them; it came so close to crushing Autrey that it left grease on his cap.

On December 11, 2008, Bernard Madoff was arrested and charged with securities fraud. For decades, Madoff – a revered member of the New York financial elite – had been running a Ponzi scheme, deceiving his clients and official investigators alike. Though the damage he wrought is difficult to assess, the total loss to investors has been estimated in the tens of billions of dollars – the largest fraud in the history of money. In the aftermath, at least two of his clients committed suicide, and many charitable organizations, his favorite marks, were forced to close.

In January 2009, New Yorkers heeding President Obama's call for a new era of responsibility donated a record 925,000 pounds of food to the *Daily News*-City Harvest feed-the-hungry campaign. Approximately one million needy residents of the city benefited from these donations, which were distributed by 600 community organizations. During the worst economic conditions since the Great Depression, and at a time when New York City in particular was suffering job losses, this display of generosity impressed and encouraged.

Why did the witnesses of Genovese's brutal attack fail to help their neighbor? Why did Autrey risk his life to save a stranger? Why did Madoff defraud charities and other investors? Why did New Yorkers succor their neighbors?

One way to answer these questions and others like them is by appeal to character traits. The witnesses lacked compassion; they were callous to the plight of the innocent even when intervening would have cost them virtually nothing. By contrast, Autrey noticed and responded instinctively to the distress of a stranger. Madoff was greedy and dishonest, manifesting a shocking inclination to deceive. Ordinary New

Yorkers were generous and humane, choosing to forgo their own material benefit in order to help those in need. Traits like callousness, courage, greed, dishonesty, generosity, and humanity are dispositions to react to circumstances in appropriate, reason-guided ways. The callous person sniffs at the suffering of others; the courageous man braves dangers to secure something valuable; the dishonest financier jumps at the opportunity to fool others for his own gain; the generous citizen keeps her weather eye out for chances to do the needy a good turn.

The fully virtuous person possesses all the virtues, and so is disposed to do the appropriate thing in all circumstances. Such a disposition has counterfactual heft: the virtuous person gives when presented with the opportunity, and she *would* give *were* she presented with a similar opportunity. This metaphysically robust property underwrites the prediction and explanation of behavior. It is therefore a presupposition of theories of virtue that moral agents have – or at least could have – counterfactual-supporting dispositions.

At first blush this presupposition is fairly uncontentious. How could one deny that people are, or at least could be, just, sincere, compassionate, chaste, considerate, trustworthy, courteous, diligent, faithful, tactful, valorous, and humble? We seem to understand ourselves and one another in terms of such character traits. Williams (1985 p. 10 n. 7) goes so far as to say that objecting to the notion of character amounts to "an objection to ethical thought itself rather than to one way of conducting it." Yet skeptics argue that situational influences swamp dispositional ones, rendering them predictively

and explanatorily impotent. And in both science and philosophy, it is but a single step from such impotence to the dustbin.

I have claimed that folk psychology goes in for character traits, but folk intuitions are notoriously fickle, and an argument can be made that they also contradict presuppositions about the existence of character traits. When the comedian John Hodgman (2001) conducted an anecdotal survey, asking people whether they would prefer the power of flight or invisibility, respondents almost uniformly characterized invisibility as the "sneaky" power. One woman said she would choose invisibility so that she could steal cashmere sweaters from department stores. A man chose invisibility, offering the following reason: "You'd have the ability to spy on people, like your exes. And that would be fun and enlightening, and a little bit perverted." He then reconsidered: "Invisibility leads you… um… leads me, as an invisible person, down a dark path. Because you're not going to want to miss out, when you're invisible – because no matter how many times you've seen a woman naked in the shower, you're going to want to see it again, because there's always a different woman." My own experience corroborates Hodgman's: by and large, when asked what they would do with the power of invisibility, men say they would peep at naked women and women say they would spy on the men in their lives to ensure their fidelity.

Are individual dispositions really so frail? Are circumstances really so powerful? The primary aim of this project is to argue that empirical challenges to moral theories like virtue ethics should be co-opted rather than resisted. Virtue ethics has much to offer. Its vision of a flourishing life seems a better object of moral contemplation and

evaluation than the desiccated, ascetic rules of deontology and consequentialism; its focus on “thick” concepts like honesty and courage seem to bridge the *is/ought* gap; its weaving together of reasons and motivations obviates concerns about moral schizophrenia. Furthermore, the virtue ethical account of action paints a detailed picture of sensitivity to reasons, careful and correct construal of ambiguous information, thoughtful deliberation, and reason-motivated action in accordance with the upshot of deliberation.

Recently, however, philosophers informed by the situationist tradition in social and experimental psychology (chief among them John Doris and Gilbert Harman) have begun to question the empirical presuppositions of virtue ethics, and a cottage industry has grown up around attacking and defending their arguments. The disputants often seem to be talking past each other, with critics of virtue ethics claiming some empirical presupposition is false or unsupported and defenders of virtue ethics responding that it was never a presupposition in the first place. To move the debate forward, I develop a comprehensive list of the empirical presuppositions of virtue ethics, the most contentious items of which are *consistency* (if an agent possesses a virtue sensitive to reason *r*, then she responds to *r* whenever it is relevant), *explanatory power* (if an agent possesses a virtue, then reference to that virtue sometimes explains her behavior), *predictive power* (if an agent possesses a virtue, then reference to that virtue sometimes enables prediction of her behavior), and *egalitarianism* (almost anyone can be virtuous).

The empirical challenge articulated by Doris and Harman targets the conjunction of these four claims. Several decades of studies in social psychology have shown that

most people respond not to the reasons there are for them to act on a given occasion but to situational factors ambient sounds, ambient smells, moods and emotions, presence of bystanders, and social distance. These morally irrelevant but causally powerful factors can be unified under the heading of attentional focus: loud and annoying sounds, unpleasant smells, negative moods and emotions, and the presence of bystanders all lead to the focusing of attention on a small number of situational features, while their opposites lead to the dilation of attention. When their attention is focused, people exhibit inattentional blindness, which leads them to miss or misconstrue important moral features of their situations; insensible to the reasons there are for them to act, they deliberate poorly (if at all) and act in violation of virtue. What explains and predicts the behavior of most people, then, is not the reasons to which they are sensitive but the situational cues that determine their attentional focus.

Situationist psychology does not just deny character traits. It also explains away the strongly felt intuition that there are character traits. It does so using a virtual pantheon of gods of error and ignorance, including the power of (mis)construal (misinterpreting ambiguous information as evidence for character traits), selection bias (using non-representative samples of behavior, thus overlook cases where people act in violation of traits), availability bias (assuming that first impressions are representative), and confirmation bias (seeking and using only evidence that confirms first impressions). These mechanisms guarantee that intuition would lead us to believe in traits even if traits did not exist. Since one does not know that *p* if one would continue to believe *p* were it false, we cannot know on the basis of intuitions that character traits exist.

Three primary responses to the situationist critique can be identified in the literature. *The dodge*: virtue is a rare ideal, so evidence that suggests most people are not virtuous is irrelevant. *The counterattack*: the empirical evidence does not support the situationist critique. *The retreat*: although the situationist critique shows that reasons-responsive traits do not exist, a naturalistic theory of virtue can still be formulated in terms of actions or local traits. Unfortunately, most versions of these arguments either are unsound or give up the consistency, explanatory power, predictive power, or egalitarianism of virtue. The dodge, for instance, is an outright denial of egalitarianism. Most versions of the counterattack fail to individuate virtues by their characteristic reasons, and thus are morally inadequate; others rely on intuitions, which are unreliable. Two compatible tactics for dealing with the challenge, however, do emerge: an emphasis on the portability of context and a shift from situation-consumerism to situation-producerism. By recognizing the power of situations and identifying the types of situations (not) conducive to behavior in accordance with virtue, one can strategically seek (avoid) situations likely to lead to (non-)virtuous actions. And by recognizing the causal dialectic between agents and situations, one can shift to thinking of agents as active producers of situations rather than mere passive consumers of them – a point of view that encourages the creation of situations conducive to action in accordance with virtue.

Along these lines, I argue that virtue (though not vice) attributions of the right sort should be made regardless of their truth-value. Drawing on formal work in multi-agent epistemic logic and empirical studies in social psychology, consumer research, and

behavioral economics, I show that the plausible, public attribution of virtuous traits induces both identification with those traits and belief that others expect one to act in trait-consonant ways, which in turn leads to trait-consonant behavior. The notions of placebo effects and self-fulfilling prophecies are instructive parallels to virtue-labeling. Thinking of virtue attributions merely as fact or fiction is too limited. We must recognize in addition a third category: factitious attributions, which make themselves true by being plausibly, publicly announced.

In another example of the portability of context and situation-producerism, I present a novel theory of social distance in terms of *potential for interaction*, *group identity*, and *information*. This theory draws support from recent work by experimental psychologists and economists, as well as an experiment that I myself conducted. Manipulating heuristics that track social distance leads agents to systematically underestimate it, which in turn leads to elevated levels of action in accordance with virtue.

I conclude by pointing to future directions for research in empirically-informed ethics. One such direction is the expansion of the situationist challenge beyond virtue ethics. I take the first step in this direction by arguing that preferences and pleasure-causation are indeterminate and troublingly dynamic, which spells trouble for many versions of consequentialism. Another direction for future work is more directly related to sections 8 and 9 of this project; it involves defending the empirical, theoretical, and moral adequacy of the notions of portability of context, situation-producerism, and factitious virtue. Finally, I make a few remarks on the implications of situationism for the

concept of the self. The bounds of the situation are the bounds of the self. If the characterization of situations I develop in sections 5 through 9 is correct, though, the self is more like an inflated balloon than a solid ball: my situation is not just my external environment, but also seemingly internal properties like my moods. The self is thus thin in a very real sense: both external and internal pressures conspire to give it its shape.

Chapter 2. The virtues of virtues

He has all the virtues I dislike and none of the vices I admire.
~ Winston Churchill, apocryphal

The clarion call for the revival of virtue ethics was Anscombe's feisty "Modern moral philosophy" (1958), in which she criticized her article's namesake. She claimed that it is not profitable to do ethics until we possess a proper philosophy of psychology – one that provides a theory of reasons, motives, and dispositions. Furthermore, she argued, the discipline of moral philosophy has been led down the garden path by the seeming generality of terms like "ought," "should," "right," and "good." These terms are too general and thin to be of any use in reasoning about actual human moral psychology, and they derive from an outmoded theory of divine command. Worse still, they shift focus from agents and whole lives to particular actions; morality thus devolves into picayune casuistry.

While one might be inclined to contest these points, Anscombe's emphatic return to Aristotle made its mark on value theory.[1] Since 1958, a sizeable contingent of moral philosophers has paid close attention to notions like *eudaimonia* (translated variously as "happiness," "flourishing," and "well-being"), character, and virtue, in addition to the

[1] Casuistry, however, was not abandoned, as evidenced by the metastasized literature on runaway trolleys.

more common contemporary focal points of goodness, rightness, and obligation.[2] Virtues became such a hot topic that even those like Rawls outside the virtue ethical tradition felt it incumbent upon themselves to address the virtues.[3] A raft of metaphysical, conceptual, methodological, and empirical arguments has been advanced for preferring virtue ethics over deontology and consequentialism:

(2.1) The proper objects of moral *contemplation* are not deeds or occurrent motives, but something broader – either behavior-producing traits or whole lives.

(2.2) The proper objects of moral *evaluation* are not deeds or occurrent motives, but something broader – either behavior-producing traits or whole lives.

(2.3) Theorizing about virtues and character transports moral discourse from the rarified air of abstract principles into the evaluatively and descriptively "thick" realm of motives and reasons.

(2.4) Reflecting on the virtues is a better guide to action than reflecting on abstract principles.

(2.5) The conceptual apparatus of virtue ethics helps one bridge the *is/ought* gap.

(2.6) Moral cultivation or education is more effective when the focus is on virtues and character than when it is on the application of abstract rules.

---

[2] See Hursthouse (1999, p. 74), Kupperman (1995, p. 7; 2001, p. 250), Oakley (1996), and Taylor (1991, p. 108).

[3] For Rawls (1971, p. 436), virtues are "strong and normally effective desires to act on the basic principles of right."

2.1. Virtue and moral contemplation

The first point can be construed as methodological. When we think about morality, we most fruitfully ask "How should one live?" rather than "What is our duty?" or "How may we be good?" or even "How can we be happy?" (Williams 1985, p. 4).[4] This dilation of focus away from the quotidian and the next moment enables one to engage in long-term projects, which are arguably necessary for a fulfilling life.[5] In his reflections on *eudaimonia*, Aristotle (1101a) makes a similar point: "the happy person is the one who, adequately furnished with external goods, engages in activities in accordance with complete virtue, **not for just any period of time but over a complete life**."[6] There is certainly something right about this argument. Flitting from moment to moment without long-term projects is no way to live. Lives should be more like novels than like collections of short stories.

2.2. Virtue and moral evaluation

Furthermore, when we approve or disapprove (and when we express these attitudes in praise or blame), the typical object is not what someone does. Instead, it is either the robust motive or trait from which the action flowed, or the person herself. On

---

[4] See Crisp (1996, pp. 1-2) and Slote (1992, pp. 3-8).

[5] See Geach (1977, p. 16), who argues that possession of the virtues is necessary to achieve our long-range purposes.

[6] A note on notation: in quotations I use italics to indicate the author's emphasis, boldface to indicate my own.

this understanding of evaluation, actions receive their moral flavor from the dispositions that cause them. A noble act is one such as the noble person would perform, and it is noble because the noble person would perform it. Dent (1975, p. 319), for instance, argues that temperance is (but justice is not) a virtue because "the temperance of acts is dependent on the temperance of men," whereas the justice of acts is not so dependent.[7]

2.3. Moral schizophrenia vs. virtue as a "thick" concept

Next, while moral theory should distinguish between right-making properties and motivating properties, allowing too great a schism between them amounts to a sort of "moral schizophrenia." The moniker was coined by Stocker (1976, p. 453), according to whom one "mark of a good life is a harmony between one's motives and one's reasons, values, justifications." This line of argument holds that even if consequentialism (or deontology) were true and even if people were somehow brought in line with its precepts, they would not be motivated to maximize utility (or act from universalizable maxims) as such. Hence, the justification or reason for their action would differ from the motive driving their action. Since virtue ethics advises one to do the right thing for the right reason, it avoids the theoretical mediation that plagues its competitors.[8] The

---

[7] This theory of the relation between virtuous acts and virtuous traits has its discontents, of course, most notably Hurka (2001, pp. 3-28; 2006) and Thomson (1996).

[8] See Annas (1993, p. 70), Audi (2001, pp. 82-83), Foot (1997; 2001, p. 9), Hudson (1980), McDowell (1979, p. 88), Wallace (1974, p. 193), and Williams (1985, p. 19).

thick/thin distinction is due to Williams (1985, p. 129), who says that thick concepts "seem to express a union of fact and value. The way these notions are applied is determined by what the world is like (for instance, by how someone has behaved), and yet, at the same time, their application usually involves a certain valuation of the situation, of person or actions." The use of such concepts bridges the gap between right-making properties and motivating properties, thereby obviating concerns about moral schizophrenia.

2.4. Virtue as action-guiding

Related arguments have been levied to show that virtue ethics enjoys other advantages over its peers. One such argument is that focusing on virtue concepts provides a better guide to action than focusing merely on good and bad, right and wrong, mandatory and impermissible. For instance, Anscombe (1958) claimed that asking whether a particular action is morally wrong may leave one dumbfounded, but that asking whether it was unjust would often make the answer "clear at once."

2.5. Virtue and the is/ought *gap*

Another argument has it that the thick concepts of virtue ethics (e.g., prudence, charity, modesty) capture intuitive moral psychology better than the "thin" concepts of consequentialism and deontology (e.g., goodness, rightness, obligation), and that these thick concepts help to bridge the supposed *is-ought* gap. Foot (2001, p. 7), for instance, points out that resolving virtue notions like *honest* into their descriptive and normative components is nigh impossible. An honest person is one who tells the truth because it's

the truth, but an honest person is good insofar as she is honest. As MacIntyre (1984, p. 199) puts it, "to identify certain actions as manifesting or failing to manifest a virtue or virtues is never only to evaluate; it is also to take the first step towards explaining why those actions rather than some others were performed. [… W]ithout allusion to the place that justice and injustice, courage and cowardice play in human life very little will be genuinely explicable." If one knows that Albert is honest, one can predict his behavior (truth-telling), but one can also praise him.

2.6. Virtue and moral education

Finally, according to proponents of virtue ethics, moral education is most effective when the teacher appeals not to abstract principles but to the thick terms just discussed. As a matter of empirical fact, exhortations to maximize happiness or do one's duty fall flat, whereas encouragements to act courageously and prudently catch the audience's attention and imagination.[9] Nussbaum (1995, p. 10) takes this argument a step further, saying that literature is an ideal medium of moral education because narratives use thick concepts to describe, explain, and evaluate behavior, leading empathetic readers to follow suit. For her, literature is a tool of moral education, not *l'art pour l'art*. Though this view may seem flat-footed and middle-brow, it has gained a degree of currency.

This one-sided presentation is intended not as a decisive argument in favor of virtue ethics but as goad to thinking that virtue ethics may have more to offer than its competitors. The question remains, though, whether virtue ethics is empirically and theoretically adequate. Perhaps it is merely the best of a bad lot. To answer that

[9] See section 8 for empirical corroboration of this argument.

question, we must identify the minimal commitments of virtue ethics, then assess them in light of evidence and arguments in both philosophy and the sciences.

Chapter 3. Groundwork for the metaphysics of virtue

Fear not the lions, for they are chained, and are placed there for trial of faith where it is, and for discovery of those that had none. Keep in the midst of the path, no hurt shall come unto thee.
~ Paul Bunyan, *The Pilgrim's Progress*

3.1. Theme and variations

Recall the story of Wesley Autry, who dove onto the train tracks to save Cameron Hollopeter. Surely this was a courageous deed. Is not Autrey living proof that there are virtuous people, that the situationist challenge to virtue ethics misses its mark?

Table that question for the moment and consider several hypothetical bystanders to Hollopeter's seizure and fall. The characterization of each of them ends with a sentence beginning, "However…." Let us assume that, but for the fact mentioned in this sentence, each hypothetical bystander would have done exactly as Autrey did.

Bravado Minkviol was also on the platform. However, he was listening to his iPod at maximum volume and did not notice Hollopeter's fall.[10]

Vivaria Boldmonk was also on the platform. However, she was in a thoroughly dejected mood; her depressed spirits made the whole world seem grey and

[10] For more on the power of ambient sensory stimuli, see Baron (1997), Baron & Thomley (1994), Boles & Haywood (1978), Cohen (1978), Cohen & Lezak (1977), Donnerstein & Wilson (1976), Geen & O'Neal (1969), Grimes (1999), Konecni (1975), Korte & Grant (1980), Korte et al. (1975), Matthews & Cannon (1975), and Page (1974).

uninteresting, so she too missed Hollopeter's fall, along with more or less everything else that happened.[11]

Dorian Vivalcomb was there too. He was in great spirits, having just won a small lottery. His bright mood led him to attend to every feature of his environment, so he immediately noticed Hollopeter's fall. However, Dorian mistook what he saw for performance art and assumed nothing was the matter.[12]

Blavdak Vinomori was also waiting on the platform. He was a visitor in New York City, unsure what to expect. When he saw Hollopeter fall, he looked around to see how others would react. However, none of them seemed concerned, leading him to believe nothing was the matter, so he went back to perusing his tourism guide.[13]

Vivian Darkbloom was waiting for the train as well. He saw Hollopeter fall. However, because Hollopeter was white and he was black, and becauseshe was convinced by Spike Lee's vile diatribes that people from historically subjugated groups

---

[11] For more on the power of bad moods, see Apsler (1975), Carlsmith & Gross (1968), Clark, & Karp (1978), Clark & Schwartz (1976), Isen, Shalker, Clark, & Karp (1978), Isen (1987), and Regan (1971).

[12] For more on the power of good moods, see Schaller & Cialdini (1990).

[13] For more on the power of bystanders, see Latané & Darley (1968, 1970), Latané & Nida (1981), Latané & Rodin (1969), and Schwartz & Gottlieb (1991).

have no social obligations to descendants of oppressors, he decided not to help Hollopeter as he lay on the tracks, his body wracked by seizures.[14]

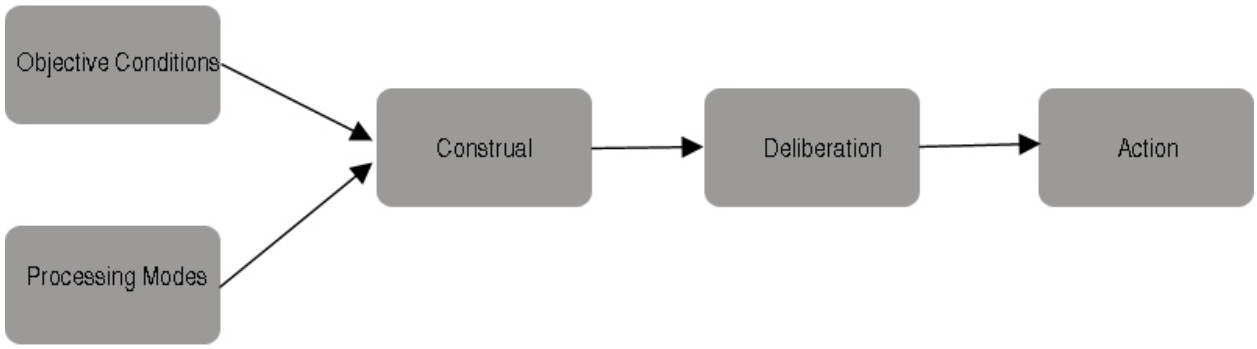
Vivian Bloodmark was yet another commuter on that eventful day. He noticed Hollopeter's fall, realized it was an emergency, and decided that the best course of action would be to rescue him. However, when he heard the oncoming train, she froze in fright.

Do Bravado Minkviol, Vivaria Boldmonk, Dorian Vivalcomb, Blavdak Vinomori, Vivian Darkbloom, and Vivian Bloodmark share anything of moral value with Wesley Autrey? Which of them, if any, possesses the virtue of courage?

3.2. Stages in the metaphysics of virtue

Consider the following plausible, albeit simplified, model of the production of behavior:

Figure 1: Model of behavior production



[14] For more on the power of social distance, see Bohnet & Frey (1999b) and Hoffman, McCabe, & Smith (1996).

To answer the questions posed at the end of the previous section, we need to decide which of the boxes in this figure fall within the scope of being virtuous. To put the same point in a different way, what we need is a principled way of deciding which stages in the production of behavior are implicated in the metaphysics of virtue, where a stage is implicated in the metaphysics of virtue just in case any adequate account of virtue must appeal to that stage. I will use existing accounts as proxies for adequate accounts, recognizing of course that there may be theories unattested in the literature and that theories in the literature may be inadequate. The point now is not to articulate the one true theory of virtue but to lay down necessary conditions for any adequate theory of virtue.

Objective conditions can be individuated in two ways: by their causal powers and by the reasons they provide. On the causal-powers model, situation S_1 differs from situation S_2 for agent *a* just in case S_1 and S_2 tend to produce different effects on *a* (either proximally on *a*'s subjective states or distally on *a*'s behavior). On the reasons-providing model, by contrast, situation S_1 differs from situation S_2 for agent *a* just in case the reasons there are for *a* differ in S_1 and S_2.[15] Clearly, the kind of traits at stake for virtue theory require objective conditions to be individuated in the second way. Being courageous, if it has anything at all to do with objective conditions, means

[15] For more on the distinction between reasons there are and reasons one has, see Williams (1981). Russell (2009) and Sreenivasan (2002, p. 59), among others, concur in individuating situations by the reasons they provide.

responding to courage-eliciting circumstances (i.e., circumstances in which there is reason to act courageously) in appropriate ways.

Processing modes like moods and affects play a different role: they influence openness to new experiences and attentional focusing, which in turn determine which aspects of the objective conditions an agent notices and finds salient.[16] Good moods induce openness to new experiences and attentional dilation, whereas bad moods induce closure to new experiences and attentional focusing. Because Vivaria Boldmonk was in poor spirits, she failed to notice Hollopeter's fall. Attentional focusing is also influenced by environmental variables like the volume of ambient sounds. Bravado Minkviol blasted his music, leading him to overlook Hollopeter's fall. Perhaps he even made a policy of listening to loud music so that he could more easily ignore schnorrers. Vivaria's and Bravado's failures to notice the emergency arguably constitute moral failures on their parts, which would implicate processing modes in the metaphysics of virtue.

Construals are the subjective counterpart of objective conditions: agent *a*'s construal is what she takes her situation to be. Dorian Vivalcomb's attention was dilated; he was open to new experiences. Hence, he noticed Hollopeter's fall. However, because he construed the situation as performance art rather than as emergency, he did nothing. Along the same lines, Blavdak Vinomori was unsure whether the objective conditions constituted an emergency. Seeing that others failed to react, he piggybacked on what he perceived to be the locals' construal of the objective

[16] For more on processing modes, see Sizer (2000).

conditions and decided the situation did not merit a reaction. The courageous person arguably possesses the perceptual and cognitive apparatus requisite for construing objective conditions *as* courage-eliciting if and only if they *are* courage-eliciting. If this is right, Dorian Vivalcomb and Blavdak Vinomori are no exemplars of courage, and construal is implicated in the metaphysics of virtue.

Deliberation, too, is arguably implicated in the metaphysics of virtue. It enables one to determine what would in fact be the virtuous response, given one's construal. Vivian Darkbloom noticed Hollopeter's fall. He construed it correctly as an emergency. However, his unsound deliberative response was that it did not provide him a reason to act. If he did not exemplify the virtue of courage (or, perhaps, of compassion or benevolence), it must be because he deliberated wrongly, which would implicate deliberation in the metaphysics of virtue.

Finally, action too is presumably implicated in the metaphysics of virtue. Someone whose processing mode allows her correctly to construe her situation as providing the reasons that it actually does provide, and who deliberates soundly about what the appropriate response would be, counts as virtuous only if she then goes on to act on the basis of her deliberation. Otherwise, he is weak-willed (like Vivian Bloodmark) or perhaps even downright vicious (recognizing the virtuous response for what it is, but spurning it).

Thus, we have prima facie arguments for the relevance of objective conditions, processing modes, construals, deliberation, and action to the metaphysics of virtue.

This generates a taxonomy of virtue theories based on which of the five they countenance, yielding thirty-two different types of virtue theory.

Table 1: Taxonomy of virtue theories

Theory Type	Objective Conditions	Processing Modes	Construals	Deliberation	Action
1	Y	Y	Y	Y	Y
2	Y	Y	Y	Y	N
3	Y	Y	Y	N	Y
4	Y	Y	Y	N	N
5	Y	Y	N	Y	Y
6	Y	Y	N	Y	N
7	Y	Y	N	N	Y
8	Y	Y	N	N	N
9	Y	N	Y	Y	Y
10	Y	N	Y	Y	N
11	Y	N	Y	N	Y
12	Y	N	Y	N	N
13	Y	N	N	Y	Y
14	Y	N	N	Y	N
15	Y	N	N	N	Y
16	Y	N	N	N	N
17	N	Y	Y	Y	Y
18	N	Y	Y	Y	N
19	N	Y	Y	N	Y
20	N	Y	Y	N	N
21	N	Y	N	Y	Y
22	N	Y	N	Y	N
23	N	Y	N	N	Y
24	N	Y	N	N	N
25	N	N	Y	Y	Y
26	N	N	Y	Y	N
27	N	N	Y	N	Y
28	N	N	Y	N	N
29	N	N	N	Y	Y
30	N	N	N	Y	N
31	N	N	N	N	Y
32	N	N	N	N	N

In this section, I make a case for the necessity of each of these five elements to the metaphysics of virtue, leaving to one side the question whether they are jointly sufficient. Along the way, I associate extant theories with various of the thirty-two possible theories taxonomized here.

3.2.1. The necessity of objective conditions

If taking objective conditions into account is necessary for virtue possession, we may eliminate theories 17-32; if it is unnecessary, we may eliminate theories 1-16. At first blush, it might seem obvious that objective conditions are implicated in the metaphysics of virtue. The virtuous person does not simply do the right thing given his construal of the situation or in light of his deliberation; he does the right think punct. Annas (1993, p. 43), drawing on Aristotle (1105b), for instance, claims that the "virtuous person is," among other things, "the person who does in fact do the morally right thing."

It may come as a surprise, then, that much recent work in virtue theory contravenes this view. Sreenivasan (2002, p. 58; see also Snow 2009, p. 100), for instance, argues that objective conditions are strictly irrelevant to the metaphysics of virtue, opting instead to rely solely on construals. While he is right that objective conditions are of less predictive value than the agent's construals, the question is not how to predict what someone will do but to define what it would mean for someone to possess a virtue. It would be a strange theory indeed that counted someone honest even if he made a habit of stealing, lying, and cheating whenever possible, comforting himself on his own moral worth by thinking precisely these actions honest. Yet Sreenivasan seems to subscribe to exactly such a theory when he argues that a virtue

has been successfully operationalized only if the subject (the agent whose possession of the trait is in question) and the observer (the theorist) agree on what would count as behavior in accordance with the virtue. If this is right, a subject committed to mendacity cannot be convicted of dishonesty by anyone with a correct conception of the virtue in question, since they would necessarily disagree on whether the lying liar lied.

In a follow-up to his (2002), Sreenivasan (2008; see also Webber 2006a, 2006b) attempts to reattach virtue to objective conditions, but in an even more dubious way. He begins by pointing out that that the meaning of the term 'cross-situationally consistent trait' is a function of the principle of situation-individuation. An agent is consistent in all situations of type *S* if and only if she reacts in the same way whenever she is in *S*. In the previous section, I argued (citing Sreenivasan 2002, in fact) that the principle of individuation relevant to virtue ethics relies on the reasons there are for someone to act. If this is right, then cross-situational consistency for virtue ethics amounts to responding in the same way to all situations that proffer the same reasons. Oddly, Sreenivasan (2008, p. 604), claims that "a trait is cross-situationally consistent if it is manifested across situations that differ in respect of the kind of feature inviting behaviour that manifests that trait," contrasting this with a principle of individuation according to which a trait is cross-situationally consistent if it is manifested "across situations in which this feature remains constant, but other features vary."

To see why both of these principles of individuation are non-starters, suppose that Autrey was in a good mood when he saved Hollopeter's life, and that he would have acted as Table 2 indicates:

Table 2: Cross-situational consistency

	Oncoming Train	**Charging Bull**	**Knife-Wielding Murderer**
Good Mood	Saves		
Neutral Mood	Doesn't save		
Bad Mood			

According to Sreenivasan's preferred principle of situation-individuation, Autrey should be considered courageous. Holding courage-irrelevant factors like mood constant, as long as he responds appropriately to all courage-eliciting circumstances, he possesses the trait. That this notion is mistaken can be demonstrated by considering Table 3:

Table 3: Cross-situational consistency again

	Oncoming Train	**Charging Bull**	**Knife-Wielding Murderer**
Good Mood	Doesn't save		
Neutral Mood	Saves		
Bad Mood			

Sreenivasan is committed by his principle of situation-individuation to saying that if Autrey would have behaved as Table 3 indicates, he should also be considered courageous. But the Autrey of Table 2 and the Autrey of Table 3 do precisely opposite things. It hardly makes sense to attribute the same character trait to both. Worse still, if cowardice is operationalized as failing to do the courage-appropriate thing in courage-eliciting circumstances, and if the Autreys of Table 2 and Table 3 are placed in courage-eliciting circumstances while in a good mood, the former is courageous and the latter cowardly, but if they are placed in courage-eliciting circumstances while in a bad or

neutral mood, the attributions are reversed: the Autrey of Table 2 is cowardly and the Autrey of Table 3 is courageous. Presumably cowardice and courage are contraries, though; the same person cannot possess both traits at the same time. Clearly, then, both trait-relevant and trait-irrelevant features of the situation should be allowed to vary if full-fledged virtues are being tested. The appropriate principle of individuation is more coarse-grained: it appeals to the reasons there.

Sreenivasan's gaffe notwithstanding, the main point should be clear: objective conditions (in particular, the reasons there are in a given situation) are a necessary part of a theory of the metaphysics of virtue.

3.2.2. The necessity of processing modes

Reconsider a morally oblivious agent like Bravado Minkviol or Vivaria Boldmonk. Minkviol never turns his music down, so he almost invariably fails to notice the reasons there are for him to act. In the few cases when he does notice that someone is in distress, that an injustice has been committed, or that it would be rude to act thusly, he construes the situation correctly, deliberates soundly, and acts in accordance with his deliberations. But these cases are exceedingly rare. It seems to me that virtue requires a sort of moral attentiveness. As pointed out above, moods, affects, emotions, and environmental distractions (among other things) influence attentional focusing and dilation. It follows that a virtuous person must have the right affective and attentional states in order to see the reasons there are for him to act.

This view of virtue is exemplified by McDowell (1979), who argues that virtue requires sensitivity to the morally salient features of one's situation. He even goes so

far as to characterize virtue as a *perceptual* capacity. Brandt (1970) and Foot (1997) also point out that the virtuous person must be sensitive to the morally relevant features of her situation. Advocates of the ethics of care have also emphasized the moral value of attentiveness. There is no small disagreement on how best to attain such attentiveness. Tronto (1993, p. 128) thinks that passive, non-conative detachment enables one to notice morally salient situational cues, whereas DesAutels (2004) argues for engaged "moral mindfulness," which actively searches for moral and social cues.

I tend to agree with DesAutels, bearing in mind Popper's (2002) point about observation. He began a lecture by asking his hard-nosed empirical audience simply to observe. Nonplussed, they cast about the room, unsure what to do. After letting them stew for a minute, Popper went on to point out that we rarely simply observe (as Tronto would have us do); in general, we observe *x* for *F*. In other words, we attend to specific parts of our environment, looking for the instantiation of particular properties. If this is right, DesAutels's model of moral mindfulness trumps Tronto's. In any case, however, processing modes belong firmly in an account of the metaphysics of virtue, allowing us to eliminate theories 9-16.

3.2.3. The necessity of construals

It has become common practice among virtue theorists to emphasize the importance of construal. Kamtekar (2004), Kupperman (2001), Russell (2009), Snow (2009), and Sreenivasan (2002, 2008), among others, point out that the virtuous person does not just do the right thing, but does it for the right reason. Such reason-guided action requires that one recognize the reasons there are for one to act; to put the same

point another way, one can only act for the reasons there are to act if one believes they are the reasons there are to act. In the vignettes I recounted at the beginning of this paper, Dorian Vivalcomb and Blavdak Vinomori both failed to construe their situations properly. The one took it to be performance art, the other an inexplicable non-emergency. Despite their sensitivity to their environment, they failed to form correct beliefs about the reasons there were in that environment. Though they did not knowingly act contrary to virtue, they failed to act in accordance with virtue because they misconstrued their situations.

Certain virtue theorists have denied the importance of construals. They tend also to deny the importance of processing modes (if they bother to think about them) and of deliberation. For instance, Driver (2001) has advanced a utilitarian virtue theory according to which being virtuous just means tending to act in such ways as to maximize the good. We may group all such theories under the heading of *pure externalist* virtue theories, in the sense that they include neither processing modes, construals, nor deliberation in the their account of virtue (expect in the secondary and attenuated sense in which a person who has appropriate internal states will more reliably produce virtuous acts). Theories according to which at least one of these elements is a necessary part of virtue should, by contrast, be considered *internalist* virtue theories. As there are (at least) three different internal features one might countenance, there are eight different varieties of internalist virtue theory (though I find it

hard to imagine anyone committing herself to a theory that countenanced processing modes without countenancing at least one of the other two).[17]

The literature teems with complaints about pure externalist virtue theories. The most common complaint about them is that, intuitively, being virtuous is not captured merely by doing what there is reason to do but doing it because there is reason to do it. Annas (1993, p. 43) puts the point in a representative way: "The virtuous person is not just the person who does in fact do the morally right thing, or even does it stably and reliably. She is the person who *understands* the principles on which she acts, and thus can explain and defend her actions." Any theory that concurs must include construals in the metaphysics of virtue, thereby eliminating theories 5-8.

3.2.4. The necessity of deliberation

While processing modes like sensitivity or moral mindfulness enable people to notice the reasons there for them to act, and construals are the beliefs they form in light of their (in)sensitivity, deliberation is the weighing of reasons to reach an all-things-considered judgment. It is far beyond the scope of this project to give a full theory of deliberation, but one I find quite plausible is Russell's specificatory theory, according to which deliberation specifies what concrete action would in fact be the right one in light of

---

[17] Note that according to these definitions a theory may be both externalist (if it countenances objective conditions or actions) and internalist (if it countenances processing modes, construals, or deliberation); for short, we can call such a theory *mixed*. A *pure* theory, by contrast, countenances only external or only internal stages in the production of behavior.

the reasons one has identified and their various weights. Although McDowell (1979) stresses the analogy between reasons-sensitivity and perception, he is in the minority in discounting the importance of deliberation. By far the most common view in the virtue ethics literature is that deliberation or practical rationality is a crucial – perhaps the crucial – element of virtue.[18]

This near-consensus is not implausible. After all, Vivian Darkbloom was sensitive to the morally salient features of her environment and correctly construed the situation as an emergency, but he failed to respond appropriately because he judged that the all-things-considered best thing to do was to watch Hollopeter be crushed by the train. Any virtue theory that endorses such a decision as virtuous is so prima facie implausible that I doubt it should be taken seriously. If this is right, deliberation too is implicated in the metaphysics of virtue, eliminating theories 3 and 4.

3.2.5. The necessity of action

The final stage in the production of behavior is action. Since the road to hell is paved with (mere) good intentions, I take it as fairly obvious that being virtuous must culminate in virtuous action. Failure to connect the internal mechanisms already discussed with action would amount to failure to exemplify virtue. Whatever else the virtuous person is, he is someone who *acts*, thus eliminating theory 2. This completes the elimination process: only theory 1, which countenances all five stages in the production of action, remains.

---

[18] See, among many others, Annas (2003), Foot (1997, 2001), Kamtekar (2004), Miller (2009), Russell (2009), Upton (2009), and Williams (1998).

3.3. From normative adequacy to empirical adequacy

If my arguments for the necessity of all five stages in the production of action are sound, then any adequate theory of virtue must countenance objective conditions, processing modes, construals, deliberation, and action. Call such a theory a *maximal mixed* theory of virtue. A question now arises: given that only maximal mixed theories of virtue are normatively adequate, what can we say of their empirical adequacy?

The recent challenge to virtue ethics by philosophers drawing on situationist social psychology attacks theories that countenance, *inter alia*, the first and last stages in the production of action. Situationist experiments use objective conditions, individuated by the reasons there are in those conditions, as independent variables and actions as dependent variables. Thus, *any externalist virtue theory stands in need of defense against the situationist critique*. One such theory (the morally most plausible one) is the maximal mixed theory. It is therefore surprising to note that defenders of virtue ethics (Annas 2003, Sreenivasan 2002, Kamtekar 2004) often cite the fact that they subscribe to the maximal mixed theory as though it made them immune to the situationist critique. They seem one and all to be committing a species of the conjunction fallacy, rating the probability of a conjunction higher than the probability of one of its conjuncts. In particular, they regard the maximal mixed theory as undiminished in likelihood despite the fact that it is a conjunction of the (potentially) discredited pure externalist theory and the pure internalist theory. While they are right that the maximal mixed theory contains internalist elements, they fail to notice that it also contains externalist elements. In order for their argumentative move to work, they

would have to subscribe to one of the eight pure internalist theories canvassed above. All eight of those theories, however are arguably normatively inadequate. This leaves virtue ethicists with an unpalatable menu of options: accepting a normatively adequate but empirically inadequate virtue theory, accepting a normatively inadequate but empirically adequate virtue theory, or mounting a serious defense of the maximal mixed virtue theory against the situationist challenge – a labor that they have yet to complete, and which constitutes the focus of the remainder of this project.

Chapter 4. Identifying the hard core of virtue ethics

The core is irrefutable by the methodological decision of its protagonists: anomalies must lead to changes only in the protective belt of auxiliary, observational hypothesis, and initial conditions.
~ Imre Lakatos, "Falsification and the Methodology of Scientific Research Programmes"

Philosophers of science like Lakatos (1995) sometimes speak of the "hard core" of a research program, designating by that term the conjunction of claims that the research program must defend at all costs. Other claims made by the researchers are considered auxiliary hypotheses intended to form a "protective belt" around the hard core. For instance, Newtonian astronomers allowed themselves to tweak the gravitational constant *G* in light of new data, but they could not allow themselves to give up the inverse-square law itself. A theory is not falsified until its hard core can no longer be protected by new auxiliary hypotheses, or until such hypotheses grow so *ad hoc* that researchers give up on the program.

To see whether the situationist challenge truly threatens virtue ethics, then, we must decide which propositions constitute its hard core. In this section I canvas ten candidate claims, of which I weaken two and relegate two to the auxiliary belt. The remaining six, along with the weakened two, constitute the hard core of virtue ethics. These ten claims are:

(4.1) *Acquirability*. It is possible for a non-virtuous individual to acquire some of the virtues, whether through conditioning, practice, learning, or some other method.

(4.2) *Stability*. If an individual possesses a virtue at time *t*, then *ceteris paribus* she will possess that virtue at a later time $t^{\wedge}$.

(4.3) *Consistency*. If an individual possesses a virtue sensitive to reason *r*, then *ceteris paribus* she will respond to *r* in most or all contexts.

(4.4) *Access*. It is possible to determine what the virtues are.

(4.5) *Normativity*. *Ceteris paribus*, it is better to possess a virtue than not, and better to possess more virtues than fewer.

(4.6) *Explanatory Power*. If an individual possesses (or does not possess) a virtue, then reference to that virtue (or its absence) will sometimes help to explain her behavior.

(4.7) *Predictive Power*. If an individual possesses (or does not possess) a virtue, then reference to that virtue (or its absence) will sometimes enable one to predict her behavior.[19]

(4.8) *Egalitarianism*. Almost anyone can be virtuous.

(4.9) *Real Saints*. There is a non-negligible cohort of saints in the human population.

(4.10) *Integration*. Possession of the virtues is positively correlated; in other words, if an individual possesses one virtue, she is more likely to possess other virtues as well.

I now discuss each of these candidates in turn.

---

[19] Note that I here diverge from Doris (1998, p. 507; 2002, p. 22), by splitting his single consistency condition into the consistency, explanatory power, and predictive power conditions, each of which I treat separately.

4.1. Acquirability

If virtues were innate and immutable states, virtue ethics would be a strange theory indeed. It makes no sense to encourage people to be or behave in a certain way if they *must* be or behave that way or *cannot* be or behave that way. So if the virtues could not be acquired, it would be senseless to recommend being virtuous or behaving virtuously. Furthermore, if *ought* really does imply *can*, and if virtue ethics is right in saying that people ought to be virtuous, it follows that they can be or become virtuous. Finally, if virtue-possession is praiseworthy and people are only (or most) legitimately praised for what they are responsible for, virtue-possession is something one can be responsible for. It seems hard to imagine, however, that one could be responsible for an innate trait.

4.2. Stability

Once acquired, virtues should be hard to lose. If Benny acts courageously one minute and rashly the next, it would hardly do to say that he really was courageous but became rash in the course of a few seconds. Virtue attribution would then be a merely *post hoc* game. If they pick out anything, virtue terms designate psychological features that are more than ephemeral.

I distinguish between stability of virtue-possession and stability of virtue-expression.[20] Though these two are tightly connected, it is possible for someone possessing the relevant trait to fail to express it in some circumstances. Aristotle

[20] Unlike Doris (1998, 2002) and Harman (1999, 2000, 2001, 2003, 2006).

(1095b) himself makes this point when he says that the virtuous person is still virtuous when asleep, and when he claims that some things are beyond human endurance, and that even the virtuous person should not be expected to hold up under such strains. Nevertheless, Aristotle argues that virtues are stable, saying that "actions done in accordance with virtues are done in a just or temperate way not merely by having some quality of their own, but rather if the agent acts […] **from a firm and unshakeable character**" (1105b). The vast majority of virtue ethicists since Aristotle have concurred.

4.3. Consistency

A related idea is that if someone possesses a virtue responsive to reasons of type *r*, she will exhibit responsiveness to all such reasons. As Dent (1975, p. 328) puts it, virtue causes appropriate behavior in "ever-various and novel situations." For instance, the generous person countenances the well-being of others as a reason to share resources beyond what he merely owes them. He tips waiters who provide adequate service, contributes to worthy charities, helps the handicapped cross streets, and engages in other acts of supererogatory giving. And he continues to countenance this reason to act regardless of virtue-irrelevant (viz. reason-irrelevant) features of his beneficiaries (age, sex, attractiveness, ethnicity, and ability to reciprocate), himself (economic preference, mood, and state of hurry), and his environment (ambient noise, ambient smells, and presence of bystanders).

The rider about irrelevant features is crucial. The generous person should not be expected to give when able in all circumstances. Aiding and abetting criminals is not generous. Donating one's last penny is foolish. Attempting to help someone who wants

to be left alone is rude. Nevertheless, if a man shared only with Angelina Jolie lookalikes who gave him come-hither looks he would be at best imperfectly generous. Since virtues are individuated by their characteristic reasons (e.g. "*x* is worth protecting" for courage, "ϕing would be excessive" for temperance, and "*a* deserves *y*" for justice), the sort of cross-situational consistency picked out by this condition is keyed to reasons.

4.4. Access

This requirement is fairly obvious. Any normative theory worth its salt should hold out at least the possibility of identifying its norms. Consequentialism, however, is often thought to stumble on the access requirement, since the computational complexity involved in determining what would maximize happiness (or goodness, or utility) is arguably beyond our ken. Similarly, it may seem difficult to decide whether a motive is universalizable – who really knows what the world would be like if all people declined to develop their talents? For virtue ethics, the access requirement comes down to knowing what the virtues are. As Aristotle (1105b) puts it, actions "are called just and temperate when they are such as the just and the temperate person would do." Knowing whether any particular action is virtuous, then, presupposes knowing what the virtuous person would have done in the circumstances, and knowing in general what

makes an action virtuous presupposes knowing both what the virtues are and what the virtuous person would do in any circumstances.[21]

4.5. Normativity

While theorists differ on whether the virtues *invariably* lead to good actions, they agree that in general having a virtue is better than not having it, having more is better than having fewer, and having all is better than having only some. According to Foot (1997, p. 3), "virtues are in general beneficial characteristics, and indeed ones that a human being needs to have, for his own sake and that of his fellows." Thomson (1997, pp. 282-284) likewise argues that virtue-possessing individuals are good for their communities, and that virtue-possession is also (though perhaps not exceptionlessly) good for the possessor herself.

4.6. Explanatory power

Many of the advantages claimed for virtue ethics in the preceding section relied on the assumption that virtues have explanatory power. In other words, in explaining why someone performs (or fails to perform) an action, it should sometimes be necessary to appeal to her possession (or non-possession) of a virtue. Recall that MacIntyre (1984, p. 199) says that much of human behavior would be "genuinely inexplicable" without appeal to virtues.

---

[21] Audi (2001), Foot (1997), and Prinz (2009) also endorse this requirement, but while Audi and Foot think it can be met, Prinz takes it to be the main stumbling block of virtue ethics.

Explanatorily powerful properties support lawlike generalizations, i.e., generalizations that are confirmed by observation of their instances and can be projected to novel observations.[22] For instance, being an emerald and being green are explanatorily powerful properties because the fact that all observed emeralds have been green confirms the generalization that all emeralds are green and supports the prediction that the next observed emerald will be green. By contrast, being near the surface of the Earth is not explanatorily powerful because the fact that all observed emeralds have been near the surface of the Earth does not confirm the generalization that all emeralds are near the surface of the Earth, nor does it support the prediction that the next observed emerald will be near the surface of the Earth.

Explanatorily powerful properties are natural kinds, and it is contemporary orthodoxy that natural kinds are metaphysically robust properties that can and should be investigated *a posteriori* (Kripke 1972). Furthermore, I follow Lewis (1986), Lipton (2004), and Salmon (1984) in thinking that explanatory power is grounded in (sometimes unknown) causal mechanisms. The best way to show that virtues have

---

[22] There is of course a large body of work in the philosophy of science on explanation, a few luminaries of which are Goodman (1965), Hempel (1966), and Salmon (1984). Yet virtually none of the work on explanation has informed the debate over situationism and character. This lapse is especially surprising in light of the fact that the same Harman (1965) who made the phrase “inference to the best explanation” the watchword of philosophy of science for decades is now one of the foremost critics (1999, 2000, 2001, 2003, 2006) of virtue ethics.

explanatory power, then, is to demonstrate their ability to cause (or prevent) behavior. The next best way is to demonstrate their correlation with behavior. Since nearly all psychology has yet to graduate to the level of demonstrating causal mechanisms, virtue ethicists and their critics should be content for now if it can be shown that virtue-possession is reliably correlated with behavior and that a plausible causal story connecting virtue-possession to behavior can be told.

Since explanatory power has been one of the primary bones of contention between situationists and defenders of virtue ethics, this requirement bears emphasis. If virtues have explanatory power, they should license statements like the following:

(4.11) Cristina told the truth because she was honest.

In general, if *a* is an agent, *V* a virtue property, and υ_a the performance by *a* of an action in accordance with *V*, the explanatory power requirement says that some statements of the following form are true:

(4.12) υ_a because *V*(*a*).

Furthermore, the explanatory power requirement says that counterfactual conditionals of the following form are true:

(4.13) $V(a) \rightarrow \upsilon_a$

That is, if *a* were virtuous, then *a* would act in accordance with virtue. It is well known that such counterfactual conditionals are stronger than their material cousins, so if the explanatory power requirement holds, statements of the following form are also true:

(4.14) $V(a) \supset \upsilon_a$

And there's the rub. Situationist critics of virtue ethics are fond of pointing out that in many circumstances where one would expect action in accordance with virtue, one is disappointed. Since so many people lack virtue, it becomes unclear whether a sufficient proportion of them even *could* possess it.

Locating explanatory power in the hard core of virtue ethics is therefore problematic. Many arguments for virtue ethics seem to be based on it, since they claim that virtue-possession motivates one to act in appropriate ways.[23] Yet *prima facie* evidence tells against this premise. I guess I'm a Popperian at heart: I like my theories like I like my coffee – dangerously strong. I therefore place explanatory power in the hard core of virtue ethics.

4.7. Predictive power

The social sciences obviously do not formulate exceptionless laws consonant with the deductive-nomological model of Hempel (1966). Economists do not predict recessions with probability 1.0. Political scientists do not forecast elections with certainty. Anthropologists do not predict the emergence of myths without trepidation, if at all. Psychologists do not predict human behavior or mental states with anything approaching the rigor of the hard sciences.

---

[23] Most virtue ethicists who commit themselves to the explanatory power requirement do so by saying that virtues are causally efficacious. See Audi (2001, pp. 82-84), Brandt (1992, p. 13), Dent (1975, p. 328), Hudson (1980), MacIntyre (1984, p. 199), Wallace (1978, p. 193), and Watson (1990, p. 451).

Nevertheless, if virtue ethics is to have explanatory power, it stands to reason that it should have predictive power as well. How much? The minimal metric is doing no worse than chance:

(4.15) $\mathrm{P}(\upsilon_a \mid V(a)) \geq \mathrm{P}(\upsilon_a)$

For example, if a randomly chosen person can be expected not to lie in a given context with probability 0.7, then an honest person can be expected not to lie in the same context with probability greater than or equal to 0.7. Still, one feels that merely not being beaten by the market is no great shakes, especially when the market is in recession. Beefing up the requirement with a strict inequality seems only trivially better:

(4.16) $\mathrm{P}(\upsilon_a \mid V(a)) > \mathrm{P}(\upsilon_a)$

Anything else, though, feels *ad hoc*. Perhaps all can agree, though, that the following is a fair requirement:

(4.17) $P(\upsilon_a \mid V(a)) >> P(\upsilon_a)$

I reject, however, the idea that virtue-possession entails high probability of acting in accordance with virtue. While it would be nice to have something like

(4.18) $\mathrm{P}(\upsilon_a \mid V(a)) \approx 1.0$

or

(4.19) $\mathrm{P}(\upsilon_a \mid V(a)) > 0.9$

or at least

(4.20) $\mathrm{P}(\upsilon_a \mid V(a)) > 0.5$

such a requirement is too strong. The proper object of concern is not absolute probability, but divergence from the base rate.

4.8. Egalitarianism

One way to insulate virtue ethics from empirical critique is to say that most people could never become virtuous. Only an elite cadre – owing to their biology, upbringing, drive, or luck – can ever become virtuous. If this is right, then psychological experiments showing that two thirds of people will comply with an authority figure in administering potentially deadly shocks to an innocent victim (Blass 1999) can be shrugged off. Such an attitude, though plausible for Plato, Aristotle, and Nietzsche, rubs our democratic ethos the wrong way. It violates Flanagan's (1991, pp. 32-37) *principle of minimal psychological realism*: "Make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behavior prescribed are possible, or are perceived to be possible, for creatures like us."

I will weaken the egalitarianism requirement in the following way: *almost anyone can behave in accordance with virtue; indeed, almost anyone can behave in accordance with virtue reliably*. I do not go so far as to require, however, that almost anyone can *be* virtuous. Thus, I demand that an adequate theory of virtue ethics make a case that almost anyone can reliably be brought to do what the virtuous person would do, but I do not take the further step of demanding that almost anyone can be a virtuous person himself.

4.9. Real saints

Another way to insulate virtue ethics is to say that there need not be real saints, or fully virtuous people. On this view, virtue ethics erects a regulative ideal of the saint, rather than identifying actual individuals to emulate. A confident virtue ethicist should

accept the real saints requirement. If *ought* implies *can* and people ought to be virtuous, then they can be virtuous. And the best evidence that something is possible is that it is actual, so a confident virtue ethicist should be willing to point out particular people as real saints.

Nevertheless, I see little point in debating the real saints requirement given the current state of psychological research restrictions. A scientific study to determine whether a given individual is a saint would require systematically tempting and tormenting the poor person. For better or worse, no institutional review board would approve of testing people in the way Job was tested, so we will have to rely on anecdotal evidence. The number and strength of confounds to such evidence is insurmountable. Hence, while I would like to endorse the real saints requirement, I see no way to test it and therefore reject it as moot.

4.10. Integration

The integration requirement is a weaker version of the well-known unity of virtue thesis, according to which someone (fully) possesses any particular virtue if and only if she possesses all virtues. To defend this *prima facie* implausible theory, one may point out that the courageous person is not rash, and so does not enter into dangerous situations without a little prudence. Integration requires less. Let V_1 and V_2 be virtue properties and *a* an agent. According to the integration requirement:

(4.21) $\mathrm{P}(V_1(a) \mid V_2(a)) > \mathrm{P}(V_1(a))$

Someone is more likely to be just given that she is also courageous, more likely to be temperate given that she is also humble, more likely to be honest given that she is

faithful. Annas (1993, p.75) commits herself to the integration thesis when she claims to have noticed that thieves are “also found telling lies.” I find the integration thesis heroic but implausible. Since it does not seem essential to virtue ethics, I therefore reject it with some relief.

Chapter 5. The empirical challenge of situationism

One is best punished for one's virtues.
~ Friedrich Nietzsche, *Beyond Good and Evil* 4.132

I have set high standards for a plausible theory of virtue ethics, requiring a maximally mixed theory that conforms to the acquirability, stability, consistency, accessibility, normativity, explanatory power, predictive power, and egalitarianism conditions. Now I turn to the broadside attack mounted by Doris (1998, 2002) and Harman (1999, 2000, 2001, 2003, 2006) against the consistency, explanatory power, predictive power, and egalitarianism requirements.[24]

To motivate their critique, I begin with a discussion of a few famous (or infamous) experiments, the first of which is Darley & Batson's (1973) Good Samaritan study, which was conducted with subjects from the Princeton Theological Seminary. Participants filled out a questionnaire to determine whether they related to religion as a means, an end, or a quest. They were then asked to prepare a talk either on job prospects for seminarians or on the New Testament parable of the Good Samaritan, in which a robbed and beaten man is ignored by a priest and a Levite but helped by a lowly Samaritan. The moral of the story of course is that one should emulate the compassionate Samaritan, not the sanctimonious clergy. Presumably the seminarians knew this, and presumably each of them wanted to follow Jesus' advice.

---

[24] Campbell (1999) and Gibbard (1992) also recognize the power of seemingly inconsequential situational factors to influence moral behavior.

To test whether the seminarians would act on the lesson they were about to teach, Darley and Batson arranged for each of them to encounter a distressed confederate slumped on the ground along the path from the questionnaire station to the speech station. Some were told that they had time to spare, others that they were just on time, and still others that they were running late. The experimenters covertly observed whether the subject stopped to help like the Good Samaritan about whom they were to sermonize. Before reading the next paragraph, ask yourself: which of the treatment conditions influenced their behavior?

If you predicted that their view of religion or the subject of their talk made the difference, you are like most people – attributing behavior to internal variables. And like most people you are mistaken. Subjects' religiosity and speech subject were uncorrelated with their helping behavior, but their degree of hurry made all the difference. Despite the fact that they were reenacting the very parable about which they were to lecture, a huge majority (90%) of the rushed participants failed to show compassion; by contrast, those without a sense of hurry helped 63% of the time. Since the treatments were randomly assigned, we must assume that the same subjects who failed to help when rushed would have helped had they been unrushed, and conversely.

Next, in a pair of famous experiments, Isen & Levin (1972) varied subjects' mood by giving them cookies or arranging for them to find a dime in the coin return of a public payphone. Those in the cookie experiment were then prompted either to help a classmate or to learn some trivia. When their moods were elevated by the unexpected cookies, participants were more willing both to help and to learn. Similarly, in the dime

experiment, subjects who found the change were much more willing to spontaneously help a confederate who "accidentally" dropped a sheaf of papers in their path. Had subjects in these experiments been benevolent or compassionate in the way required by virtue ethics, most would have done the same thing: helped unless they had a strong overriding reason not to. What we find instead is that their mood largely determined whether they helped.[25]

The situationist critique of virtue ethics proceeds by pointing out that if people are unconsciously susceptible to such minor influences as their degree of hurry, receiving cookies, and finding dimes, one can only infer that they would also be swayed by major temptations. As Doris (1998) puts it, "Aristotelian virtues are robust, or substantially resistant to contrary situational pressures, in their behavioral manifestations." Situationist experiments suggest that most people do not even have flimsy traits, let alone robust ones. Experiments like the Milgram (1974) obedience study and the Haney, Banks, & Zimbardo (1973) prison simulation elicited such appalling behavior in response to insubstantial situational pressures that Doris & Stich (2005) feel certain in saying that people do not merely "fall short of ideals of virtue and fortitude, but that they can be *readily* induced to *radically* fail such ideals."

Doris (1998, 2002), Doris & Stich (2005), and Flanagan (1991) largely converge on the view that while the stability requirement can be met, the consistency, explanatory

---

[25] The experiments also showed that the subjects were not vicious: they did not all fail to help. This result, however, is in perfect harmony with the situationist message that people lack character traits – good, bad, and indifferent.

power, predictive power, and egalitarianism requirements cannot be simultaneously satisfied. People can be expected to behave the same way in iterations of the same situation. If Zena contributes money to a good cause when the sun shines on Monday, she will most likely contribute to the same cause when the sun shines on Friday. People do not exhibit cross-situational consistency, however, even when the difference in situation is minor and morally irrelevant. Their future behavior cannot be predicted with sufficient certainty on the basis of previous behavior. Zena might be a fair-weather benefactor, failing to give when the sun hides its face. Doris would say that Zena lacks the global trait of charity but possesses the local trait of sunny-weather charity. While there may be a few people who exemplify the traits lionized by virtue ethics, the dearth of longitudinal studies makes any such supposition a hazardous guess, and there is no reason to think that others could follow in the footsteps of such saints even if they do exist. Empirical results cannot show that virtue is an unachievable ideal, but as Doris & Stich (2005) have argued, "the burden of argument has importantly shifted: The advocate of virtue ethics can no longer simply assume that virtue is psychologically possible."

Harman (1999) goes further in attacking virtue ethics, arguing both that "there is no such thing as character, no ordinary character traits of the sort people think there are, none of the usual moral virtues and vices," and that the local traits countenanced by Doris do not count as traits at all, thereby challenging the stability requirement as well.[26]

---

[26] There is one peculiar type of trait that would be very hard for situationism to challenge; in fact, situationism seem to *confirm* its existence. The trait in question is the

disposition to engage in consistently inconsistent behavior: being *labile, flaky, flighty, inconsistent, fickle, changeable, inconstant, mercurial, temperamental, unpredictable, unsteady, erratic, vacillating, wavering, undependable,* or *unreliable*. These traits are sometimes used to explain and even predict behavior, as the following exchanges illustrate:

Why did James have such a bad time in Hawaii, despite the fact that he saved for three years to afford the flight and hotel?

Because he's fickle.

Will Emily show up to her own birthday party?

I wouldn't count on it; she's flaky.

When situationism claims that most people are consistently inconsistent, it actually argues that such traits are widespread. I suppose one can countenance such traits without giving quarter or comfort to defenders of Aristotelian virtue ethics, since all the traditional virtues are dispositions to engage in consistently consistent behavior. Doris and Harman (personal communication) agree, though the former believes that even such inconsistent traits might be shown not to exist.

For the sake of argument, we can operationalize dependability as the disposition to arrive on time to scheduled events, respond promptly to correspondence from people who deserve replies, keep promises, and the like. Flakiness can then be defined as the contradictory of dependability: someone is flaky just in case she is not dependable. I would resist a stronger characterization. For instance, someone who is always late, never returns correspondence, never keeps promises, and so on isn't just flaky – she's

This prompts the question whether virtues and other traits are the sorts of things to be identified a priori or a posteriori. Harman's argument seems to presuppose that since there are no traits "of the sort people think there are" there simply are no traits at all. In a later article (2003), he points out the difference between articulating the commitments of a theory like virtue ethics on the one hand and (dis)confirming such commitments on the other. The contrast again suggests that the identification of virtues is a matter of armchair reflection and conceptual analysis, not fieldwork. Doris, Stich, and Flanagan, by contrast, seem to believe that traits are the sorts of thing that can be discovered a

---

consistently deceptive and callous. I take it that the unguarded situationist would say that most people are not dependable. That entails, though, that most people are flaky.

It might turn out that they're flaky in different ways – some are punctual when they scheduled the event themselves, others when the organizer of the event has the power to sanction them; some reply to emails, others to phone calls; and so on. All dependable people are alike; each flaky person is flaky in his own way.

If this is right, there are indeed real traits, namely, dispositions to consistently inconsistent behavior. That prompts the question whether there are any virtues that fall into this category. Certainly none of the traditional virtues do, but perhaps a Salvador Dalí of virtue ethics will come along and invent some. As an initial attempt, I refer to a "virtue" of poker players – unpredictability. If life is sufficiently like poker, unpredictability would indeed be a virtue.

posteriori.[27] If they are right, and if it turns out that people really do or could have local traits, then Harman (2001) is wrong to assert that what "a person with a seemingly ideal moral character will do in a particular situation is pretty much what anyone else will do in exactly that situation, allowing for random variation." Instead, if that person has the relevant local virtue, she will most likely act in accordance with it regardless of what others do.

Helping behavior, of course, is not the only possible manifestation of virtue, though one would expect it from people who are generous, compassionate, altruistic, benevolent, considerate, courteous, friendly, or humane. The difficulty with testing other virtues is that they are harder to operationalize. How does one experimentally test subjects' courage without violating their rights?

Honesty has been tested, and it too has been found to be highly sensitive to seemingly inconsequential situational variables. In one of the largest (n = 10,865) studies of character ever performed, Hartshorne & May (1928) tested students' propensity to cheat (as exhibited on in-class tests, homework, athletic contests, parlor games), steal (as exhibited in play situations and classroom situations), and lie (either to escape disapproval or to gain approval).

A naïve believer in traits would expect the students to fall into one of three groups: inveterate deceivers, middling opportunists, and upstanding truth-tellers.

---

[27] For example, Doris & Stich (2005) characterize ethical naturalism, to which they are adherents, in the following way: "*ethical theorizing should be an (in part) a posteriori inquiry richly informed by relevant empirical considerations*."

Inveterate deceivers have the consistent vice of dishonesty; they deceive whenever it would benefit them. Opportunists deceive only occasionally, when the benefit is great and the potential to be caught minimal. Upstanding truth-tellers never deceive, or only when some other ethical concern overrides the impetus not to deceive. That is what one might expect. Hartshorne and May found, however, a mean intercorrelation between different pairs of situations presenting opportunities for deception or honesty of only 0.23. They summarized their results in the following way: "[N]either deceit nor its opposite, "honesty," are unified character traits, but rather specific functions of life situations. Most children will deceive in certain situations and not in others. Lying, cheating, and stealing as measured by the test situations used in these studies are only very loosely related. Even cheating in the classroom is rather highly specific, for a child may cheat on an arithmetic test and not on a spelling test." Children who cheated on one spelling test were likely to cheat on another, but their cheating on a spelling test was only weakly correlated with other cheating behavior, and even less correlated with other types of dishonesty like stealing and lying. This local consistency supports Doris's (2002) argument that attributions of general traits like honesty are bound to fail, while attribution of local traits like in-class-spelling-test honesty may succeed.

More recently, Ariely (2008) conducted a series of honesty experiments with Harvard and MIT students. Participants at Harvard took a trivia quiz and were rewarded monetarily for each correct answer. Ariely enabled cheating for some of the students by allowing them to grade their own tests, rather than having the proctor grade them. Like Gyges with his ring of invisibility, their honest behavior ceased as soon as they were not

watched. On average, students in the experimental condition claimed to have correctly answered 10% more questions than the control group. Participants at MIT took a math quiz, but this time there were three experimental conditions. Some were asked before the quiz to write down the names of ten books they had read in high school, others to write down as many of the Ten Commandments as they could remember, and the rest to write, "I understand that this study falls under the MIT honor system." Those in the first experimental condition claimed to have solved 33% more questions than those in the control group. To the experimenters' surprise, neither of the other experimental conditions differed significantly from the other – despite the fact that almost no one could recall all ten commandments and despite the fact that MIT has no honor system. Such results are of course grist for Doris's mill: people are not honest, but they may be Ten Commandments-primed honest and non-existent-honor-code-primed honest.

In another study, Ariely asked MIT students to solve as many equations as they could in a short time. There were two experimental conditions in which cheating was possible. In the first, like the previous study, subjects graded themselves. These students claimed to have solved 77% more equations than those in the control group. In the second experimental treatment, not only did participants grade themselves, but instead of being paid directly with cash they were first paid with non-monetary tokens, which they exchanged for cash. These students claimed to have solved 184% more equations than the control – more than double the cheating found in the first experimental condition. Are people nonmonetary-currency dishonest? Ariely's studies suggest that the answer is yes.

Although the litany of situational variables – degree of hurry, cookies, dimes, spelling tests, the Ten Commandments, imaginary honor codes, nonmonetary currency – seems unmanageably diverse, a careful review of the literature suggests a way of unifying and systematizing many situational influences on behavior. Some of the primary factors are: *ambient smells, ambient sounds, mood and emotion, empathy, bystanders,* and *social distance*. In fact, I shall argue that smells, sounds, mood and emotion, and empathy can be unified under the broader headings of *attentional focusing* and *openness to new experience*. While *culture and gender* have been proposed as further systematic situational factors by Prinz (2009), the evidence is inconclusive.

5.1. Ambient sounds

The volume of ambient sounds influences both helping and aggressive behavior. People subjected to sounds at high volume (> 80 dB) consistently help less – both in emergency and non-emergency situations – than those subjected to sounds at low volume (< 80 dB). For example, Matthews & Cannon (1975; see also Boles & Haywood 1978, Korte et al. 1975, and Page 1974) found that fewer subjects were willing to help a confederate who "accidentally" dropped a belonging when background noise levels were at 85 dB than when they were at 65 dB. High ambient sound levels seem to cause attentional focusing: people attend only to the one or two most salient features of their environment. This entails an overemphasis on the focal points and an underemphasis on all else. Focusing has been successfully used to explain the effects of ambient noise both on helping behavior and on other tasks, such as noticing unusual elements of one's surroundings (Cohen 1978, Cohen & Lezak 1977, Korte & Grant 1980). This is a

phenomenon understood by iPod users in cities; an easy way to ignore beggars is to turn up the volume so that one does not even notice them.

High levels of ambient noise have also been tied to aggressive behavior. While loud noises are generally insufficient to cause aggression, people already disposed to aggress do so more frequently and more violently in the presence of high-volume sounds. Donnerstein & Wilson (1976; see also Geen & O'Neal 1969 and Konecni 1975) found that subjects given a chance to electrocute a confederate who had angered them did so more often and with higher voltages when ambient noise was loud than when it was soft. Noise had no effect on subjects who were not angered and thus not disposed to aggress. Attentional constriction helps to explain this behavior as well: when angered in high-volume contexts, people focus their attention only on the object of their rage, leading to more aggressive behavior.

5.2. Ambient smells

Congruent results have been found relating ambient smells to helping behavior.[28] Baron (1997; see Grimes 1999) showed that people engage in more helping behavior when exposed to pleasant smells than to no smells. Experimenters solicited help from passers-by in a mall. In one condition, they asked for help in front of a bakery or coffee

[28] To my knowledge, the effect of ambient smells on other types of morally important behavior has not been investigated. We do not know whether acrid smells induce courageous actions or putrid smells induce unjust deeds.

shop; in the other, they requested help in front of a dry goods store. Subjects in the first condition were more likely to help than those in the second.

Why do pleasant smells have this effect? One explanation is that they directly cause a dilation of attentional focus. Baron & Thomley (1994) conducted an experiment to investigate the effects of pleasant fragrances and direct mood elevation on both task performance and helping behavior. They found that both exposure to pleasant fragrances and receipt of a small gift (the mood elevator) increased performance on an anagrams task. In addition, both factors increased participants' willingness to volunteer. Another potential explanation of the connection between pleasant odors and helping behavior is that pleasant smells cause positive moods, which are independently connected to dilated attentional focus (as I discuss below). In either case, the end of the causal chain is broader attentional focus.

5.3. Mood and emotion

While moods and emotions are not situational in the same way that ambient smells and sounds are, they form a valid part of the situationist critique because they are *morally irrelevant* and because they are *easily induced* by trivial situational factors. A number of studies have connected mood and emotion with helping behavior.[29] Emotions and moods are difficult to tease apart, but arguably emotions have intentional

[29] See Isen, Clark, & Schwartz (1976) and Isen, Shalker, Clark, & Karp (1978). Isen (1987) reviews much of this literature.

content (either propositional or sub-propositional) whereas moods do not.[30] Emotions both positive and negative have been shown to lead to heightened helping behavior. Apsler (1975), for instance, found that embarrassed subjects were more willing to help than unabashed ones. Carlsmith & Gross (1968) used guilt to induce extra help, as did Regan (1971), who found that subjects made to feel responsible for a mishap in the lab were more willing to help in a seemingly unrelated task.

The question of emotional influence on preference has been addressed in the economics literature as well. Loewenstein (2000), for instance, points out that emotions lead to preference-dynamism, contrary to classical decision- and game-theory. This dynamism, which can be captured in state-dependent preference models, is not the same phenomenon as unpredictability or uncertainty. Rather, the effects of emotions and moods are systematic, including overvaluing the object of emotion, as well as overvaluing the present and immediate future.

Weyant (1978) complicated the picture somewhat by showing that the effect of mood interacts with both the cost to oneself and perceived benefit to others of helping. His $3 \times 2 \times 2$ analysis is summarized in Table 4.

[30] Though see Shargel's (unpublished) provocative defense of the thesis that emotions are mere bodily states with neither propositional nor qualitative content, as well as my (2010, forthcoming) less extreme theory of the tenacity of intentional states.

Table 4: Effect of mood, cost, and benefit on helping behavior

		Positive mood	**Neutral mood**	**Negative mood**
High benefit	*High cost*	Elevated	Baseline	Baseline
	Low cost			Hyper-elevated
Low benefit	*High cost*			Depressed
	Low Cost			Baseline

Note that while positive mood induced elevated levels of helping regardless of benefit and cost, negative mood was sensitive to both benefit and cost. This supports the attentional focusing and openness to experience hypotheses. Those in a positive mood are open to new experience and fail to focus on the cost-benefit ratio. By contrast, those in a negative mood are averse to new experiences and focus carefully on a single factor, the cost-benefit ratio; thus, when it is high their helping behavior is depressed below baseline, but when it is low their helping behavior is elevated even above that of subjects in a positive mood. Schaller & Cialdini (1990) argue that happy individuals differ from sad ones in their:

1. "*Access to affect-related memories and thoughts*, with positive recollections, attributions, and cognitions more accessible to elated than to saddened persons.
2. "*Attention/orientation to the environment*, with elated individuals showing a greater tendency to attend to and make contact with external stimuli in a broad, encompassing fashion, and saddened individuals showing a greater tendency to attend to the self and to make contact with the environment only in a most selective manner.

3. "*Level of arousal and physical activity*, with happy persons demonstrating greater than normal levels of physiological arousal and behavioral activity, and saddened persons demonstrating reduced arousal and activity.
4. "*Process of decision making*, with elation leading to a less controlled, more heuristic decision style, and sadness leading to a more controlled, highly considered approach.
5. "*General motivational drive*, with elated individuals striving for a counterhomeostatic enhancement of higher-order goals (e.g., affiliation, attachment, competence, achievement and saddened individuals striving primarily for the homeostatic restoration of affective equilibrium."

All of these differences, especially 2, corroborate the thesis that good mood does not lead directly to helping behavior, but rather enables one to notice opportunities to help and encourages one to undertake new projects, including but not limited to altruistic ones.

Good mood and sexual arousal both seem to lead to openness to new experiences. Ariely (2008) induced sexual arousal in 20-year-old male subjects by asking them to masturbate while watching pornographic videos. While in a "hot" state, they filled out a questionnaire, the results of which were compared with questionnaires completed by individuals in a "cold" state. Questions covered subjects' willingness to engage in a variety of sexual activities, most of which might be considered disgusting, illegal, or even immoral. Questions included the following:

(5.1) Can you imagine being attracted to a 12-year-old girl?

(5.2) Could it be fun to have sex with someone who was extremely fat?

(5.3) Could you enjoy sex with someone you hated?

(5.4) Would it be fun to watch an attractive woman urinating?

(5.5) Can you imagine getting sexually excited by contact with an animal?

(5.6) Would you tell a woman that you loved her to increase the chance that she would have sex with you?

(5.7) Would you slip a woman a drug to increase the chance that she would have sex with you?

Men in the "hot" state were uniformly more willing to engage in these activities, often by huge margins. Thus, while openness to new experience may lead to increased helping behavior, it may also lead to increased behavior of a morally problematic kind. Fair moods do not make us fair. Nor do foul moods make us foul. Instead, positive moods and emotions induce dilated attentional focus and openness to new experiences. Negative moods and emotions induce constricted attentional focus and avoidance of new experiences. Whether moods and emotions effect good or bad behavior, then, depends on what we focus on and what new experiences lie open to us.

5.4. Empathy

Empathy has been classed variously among the moods, the emotions, and even the virtues. I should think that it is none of these, but rather a tendency to enter mental states roughly congruent to those one perceives, believes, or imagines others are in. In any event, empathically induced emotion has been shown to induce helping behavior in

a variety of situations.[31] Batson (1991; see also 2002) argues for the empathy-altruism hypothesis, according to which "empathic emotion (an other-oriented emotional response congruent with the perceived welfare of another individual) evokes altruistic motivation (a motivational state with the ultimate goal of increasing the other's welfare)." The idea is that, when Nora empathizes with Oscar, she has an incentive to act as if she regards his utility as her own. Since she feels (perhaps in an attenuated, approximate way) whatever he feels, she does best if she also sees to it that he feels good. Spinoza (1677/1992) dubbed the self-destructive behavior resulting from the failure to recognize this phenomenon *human bondage*. To avoid this bondage, Nora could break her empathic connection with Oscar, thereby avoiding the ill-effects of empathizing with an unfortunate. Batson's (1991, 2002) studies show that, like the iPod users mentioned above, people sometimes do break their connection with others in just such situations, corroborating the empathy-altruism hypothesis.[32]

Since almost any mental state can be the object of empathy, however, one wants to know whether empathy as such influences helping behavior, or merely empathically induced emotions. The answer to this question is unclear, so I prefer not to count empathy as a determinant of situation in its own right. Instead, I revert to the factors of attentional focus and openness to new experiences.

---

[31] See Batson, van Lange, Ahmad, & Lishner (2003) for a comprehensive if partisan state of the art essay.

[32] Goldman (1993) and Miller (2009) also discuss empathy in this vein.

5.5. Attentional focus and openness to new experiences

Attention is an important but neglected topic in moral psychology. As evidence from the social sciences indicates, people's values, perceptions, and objective situations alone do not constitute a full explanatory base for the explanation of behavior, including moral behavior. Diverse situational factors cause our attention to constrict or dilate and our willingness to engage in new experiences to decrease and increase. Broadened focus enables us to attend to more stimuli, while constricted focus allows us to look intensely at relevant (or irrelevant) features and ignore irrelevant (or relevant) ones. Mack & Rock's (2000) study of the power of inattentional blindness emphasizes this point. They point to the now-famous study in which participants watched a video of six people passing a basketball. Asked to count the number of passes, most participants failed entirely to notice that around two-thirds of the way into the short film a man in a gorilla suit walks on to the screen, turns to face the camera, beats his chest in a threat display, and ambles off. Since they were paying attention only to the number of passes, they missed entirely this seemingly conspicuous element of the film. McDowell's (1979) view that virtue is largely a perceptual matter receives support from unexpected empirical quarters.

Above, I discussed the emblematic Isen & Levin (1972) dime study, which found subjects were significantly more willing both to help and to learn general-interest information after finding a dime in a payphone's coin return. Batson, Coke, Chard, Smith & Taliaferro (1979) replicated this result. They hypothesize that good (bad) mood is not especially moral (immoral), but that it makes people more (less) willing to engage

in all kinds of new behaviors. The Darley & Batson (1973) Good Samaritan study corroborates this speculation: presumably subjects in the hurried condition not only would have failed to help the distressed confederate but also would have declined to learn general-purpose information. They were less open to new experiences of all sorts – selfless and selfish – than were their unhurried counterparts.

Quite recently, Krupka & Weber (2006) conducted an experiment to investigate the interaction of social norms and attentional focusing with pro-social behavior. Eliminating strategic influences by having participants play a one-shot economic game, they showed that "thinking about or observing the behavior of others produces increased pro-social behavior." If people do not focus on others in this way, pro-social norms they themselves countenance are not triggered and hence do not lead to behavior they themselves would prefer. Brañas-Garza (2007) corroborated these results in a dictator game study. The control condition was a standard dictator game. The experimental condition differed only in one respect; dictators' instructions included the sentence, "Note that the recipient is in your hands." Simply drawing dictators' attention to this fact induced greater allocations to the recipient. It seems, in a way, that Socrates was doubly right: evil is committed and good omitted not so much out of ill will but out of ignorance of and lack of attention to relevant cues.

5.6. Bystanders

After the Kitty Genovese rape and murder mentioned above, a slew of social scientists began to theorize about and experiment with the so-called unresponsive bystander effect. In situations where it is common knowledge that multiple people may

intervene, helpful responses to emergencies are less frequent than in situations where only one person can intervene (or where several may intervene but none knows about the others). From a naïve dispositional point of view this phenomenon is quite strange. One would expect the probability that someone helps to increase monotonically with the number of potential helpers:

(5.8) $P(n \text{ people help}) \leq P(n+1 \text{ people help})$

In fact, however, it seems to *decrease* monotonically with the number of potential helpers:

(5.9) $P(n \text{ people help}) \geq P(n+1 \text{ people help})$

Latané & Darley (1970) hypothesized that the presence of other potential helpers reduces the probability that at least one of them comes to the victim's aid for two reasons. First, the presence of others leads to a "diffusion of responsibility." Each person feels only partially responsible for what happens because he knows that others could intervene instead, and knows that others know that he knows they could intervene, and knows that others know that he knows that they know that he knows they could intervene, and so on.[33]

Without further argument, however, the diffusion of responsibility might be expected simply to cancel any increase in helping behavior as the number of bystanders increases. But the unfortunate fact remains that helping is not constant but

[33] See Pacuit, Parikh, & Cogan (2006) for a social software approach to the diffusion of responsibility. For a definition and programmatic discussion of social software, see Parikh (2002).

monotonically decreasing with the number of potential helpers. Latané & Darley's second hypothesized cause of decreased helping may enable us to save the phenomena. They argued that in addition to the diffusion of responsibility, when bystanders are able to observe each other they rely on what they perceive others to think in construing ambiguous stimuli. As I discuss in more detail below, theorists neglect the power of construal at their peril. When one sees that others do not intervene, one tends to assume that they have disambiguating information that the situation in fact is no emergency. Everyone mistakenly takes everyone else's inaction as expressing knowledge that action is unnecessary, so everyone concludes that action is unnecessary.

Experiments have borne out these two explanatory factors. Latané & Darley (1968) found that 75% of solitary bystanders in a simulated fire emergency intervened, while only 10% intervened when two impassive confederates were present. Latané & Rodin (1969) similarly found that 70% of solitary bystanders intervened when they heard what sounded like a bookshelf collapsing on someone in the adjacent room, whereas only 7% helped when a phlegmatic confederate sat beside them. And in a study conducted by Darley & Latané (1968), subjects who heard a confederate in another room endure a simulated epileptic seizure were most likely to intervene when alone (85%), next most likely when there was one other bystander (62%), and least likely when there were four other bystanders (31%). What's more, reaction times were inversely correlated with the number of bystanders. These results are extremely robust, as Latané & Nida's (1981) literature review shows; moreover, they have stood the test

of time, as Schwartz & Gottlieb's (1991) more recent study indicates. The more people who might help, the lower the probability that anyone will help and the longer one has to wait for help.

A third potential explanatory factor for the unresponsive bystander effect once again invokes attentional focusing. People in groups tend to look at the floor and avoid eye contact, but people who (think they) are alone let their eyes and attention wander over everything in their vicinity. In group contexts, therefore, people are less likely to notice cues of trouble because they do not attend to as many cues as they otherwise would.

5.7. Social distance

Since I devote most of section 9 to social distance, I merely mention it here. The interested reader can jump directly to section 9 for a detailed investigation of the power of social distance.

5.8. Culture and gender

In a recent article, Prinz (2009) made a case for including culture and gender as situational variables influencing both acts of compassion and giving behavior. He cites cross-cultural replications of the Milgram experiment where obedience rates differed from the baseline established by Milgram (1974) of 65%. For instance, in Germany, 85% of subjects were fully obedient (Mantell 1971). In Australia, by contrast, overall compliance was a mere 28% (Kilham & Mann 1974), with women showing significantly greater defiance than men. This leads Prinz to suggest that both *gender* and *national*

character are causal determinants of behavior; for instance, Americans are hyperindividualistic, Germans extremely obedient, and Australians resolutely anti-authoritarian. He may be right, but his argument is dubious, as Blass's (1999) review of all replications of the Milgram paradigm shows. What we find is that subjects in the United States, South Africa, Jordan, Spain, Italy, West Germany, and Austria all obey at roughly the same rate. The sole exception is the Kilham & Mann (1974) study in Australia. Instead of spinning off just-so stories about national character, then, the appropriate response is to find a design feature of the Kilham & Mann study that explains why it is an outlier. This study involved a modification of the original paradigm to make it easier for subjects to disobey: the experimenter was replaced by a second confederate who was assumed to be a mere participant in the study. It was also the sole study in which a difference between genders can be detected; in all nine of the other studies that compared men and women, no gender differences emerged. Women cannot be expected to defy authority more reliably than men. Ring, Wallston, & Corey (1970) conducted a Milgram-style experiment with 57 female subjects and had a 91% obedience rate – significantly higher than the obedience rate for studies on men. Sheridan & King (1972) ran a similar study in which the learner was replaced by an adorable puppy that yelped and writhed in pain every time it was shocked (unlike Milgram's confederate, the puppy actually had electricity flowing through it). *All* female subjects obeyed to the bitter end.

Prinz (2009) also cites cross-cultural data on helping behavior, most of which uses the Isen & Levin (1972) paradigm. Here Prinz may have scored a point, but the

evidence is far from dispositive. Buchan, Johnson, & Croson (2006) and Levine, Norenzayan, & Philbrick (2001) found significant differences in spontaneous non-emergency helping behavior in major cities of different countries. However, both studies suffer from economic confounds. The fact that denizens of Rio de Janeiro were three times as likely as New Yorkers to pick up a pen for a stranger may owe more to the fact that a pen is roughly 7.5 times as valuable in Rio de Janeiro (mean income $18,106) as in New York City (mean income $135,466).

Experiments in behavioral economics have largely failed to bear out Prinz's notion that some nationalities are more generous than others. Camerer & Thaler (1995) found negligible differences in giving behavior by participants in the United States, Israel, Slovenia, and Tokyo; Charness, Haruvy, & Sonsino (2007) corroborated these results with participants in Israel, Spain, Texas, and California. Bohnet & Frey's (1999a) Swiss subjects gave about as much as most American subjects. W should not be too sanguine about finding important differences in helping behavior across either cultures or genders.

5.9. The Mischellian consensus

While I have undertaken a more extensive literature review than all other philosophical commentators on the empirical evidence, the social psychology and economics journals teem with studies I have not cited. It would take several lifetimes to discuss them all, so I hope merely to dramatize and systematize the literature on situational determinants of moral behavior. Ever since Mischel's (1968) landmark literature review, in which he showed that in nearly every objective study of personality

individual dispositions account for at most 30% of behavior (and usually between 10% and 20%), psychologists have taken a dim view of traitology.[34] Mischel further showed that though cross-situational consistency is not to be had, intra-situational stability does exist, often accounting for at least 40% of behavior in iterations of the same situation. As Allport (1966), himself a friend of traits, aptly put it, "Every parent knows that an offspring may be a hellion at home and an angel when he goes visiting. A businessman may be hardheaded in the office and a mere marshmallow in the hands of his pretty daughter."[35]

With such low correlations, the explanatory and predictive power of trait attributions are severely impugned. Even Epstein (1983), a defender of personality theory, agrees with situationists that predicting particular behaviors in particular situations on the basis of trait variables is "usually hopeless." Traits, according to him, are better evaluated by reference to general behavioral trends than particular behaviors. This is the so-called aggregation solution, according to which the proper variables to correlate are traits and behavioral patterns, not traits and individual actions. The

---

[34] Higher correlations are to be had if one also considers paper-and-pencil questionnaires designed to elicit intuitive responses, but I am skeptical of the deliverances of introspection in general, and especially skeptical when those deliverances concern subjects' character. Mischel's results for objectively measured correlations bear out this skepticism.

[35] See Mischel & Peake (1982) for a more recent literature review that comes to the same conclusion, even about conscientiousness, one of the so-called Big Five Traits.

aggregation solution, however, admits of at best a Pyrrhic victory for defenders of personality. After all, the explanatory power and predictive power requirements apply not just to trends but to individual actions. Virtue ethicists like MacIntyre (1997) think that particular human actions would be "genuinely inexplicable" and, presumably, genuinely unpredictable without appeal to traits, yet correlations of 0.30 are generally undetectable by the layperson (Jennings, Amabile, & Ross 1982).

These considerations spell trouble for virtue ethics. If correlations between virtue-eliciting conditions and behavior in accordance with virtue are below the threshold of observation, it becomes hard to understand why people would ever have attributed traits in the first place, let alone why they would continue to attribute them in the face of experience. In addition, such low correlations cast doubt on the consistency, explanatory power, and predictive power requirements. One can dodge the problem for consistency by saying that most people lack virtue, but such a move undermines the egalitarianism requirement.

Chapter 6. Explaining away intuitions about traits

Isn't it pretty to think so?
~ Ernest Hemingway, *The Sun Also Rises*

Situationism is an error theory, and error theories owe us an account of why we fall into error. Why do we have so many trait terms and feel so comfortable navigating the language of traits if actual correlations between traits and individual actions are undetectable? To answer this question, situationists invoke a veritable pantheon of gods of ignorance and error: the power of construal, selection bias and arbitrary coherence, availability bias and availability cascade, the fundamental attribution error, the false consensus effect or egocentric attribution bias, the base rate fallacy, anchoring and disregard of regression to the mean, and confirmation bias.[36] What follows is a brief discussion of each and their relevance to explaining away intuitions about the existence and robustness of traits like virtues.

6.1. The power of construal

Mischel & Shoda (1995, p.258; see also Ross & Nisbett 1991, pp. 59-89) argue that people's subjective construals of their situations account for a lot of variability in

[36] Only three of these phenomena have been addressed in the related philosophical literature to date. Somewhat oddly, defenders of virtue ethics (e.g., Sreenivasan 2002) often invoke the power of construal, but I hope to show that they are mistaken to do so. Harman has briefly discussed the fundamental attribution error (1999, p. 325) and confirmation bias (1999, p. 325).

behavior. Ambiguous environmental cues require interpretation. Was John's laugh light-hearted or sadistic? Is that person running down the street panicked or just late for a meeting? Are those cries coming from the apartment next door a plea for help from a battered wife or just the television blaring? What one person sees as an emergency calling for immediate action, another sees as a nuisance or at least as unclear.

The power of construal is relevant to the dispute over traits in two ways. First, if someone attributes one trait (say, helpfulness) to Jack and another (say, thirst for recognition) to Jill, he will interpret the same objective behavior (going up the hill to fetch a pail of water) differently depending on which person does it. Jack is trying to help out, but Jill just wants to be praised. Jill could not care less about our welfare, but Jack wants to make sure we stay hydrated. Once a trait has been attributed, all ambiguous evidence is interpreted as if it flowed from the trait.

Second, studies of the existence and robustness of traits use trait-eliciting conditions as independent variable and trait-expressing behavior as dependent variable. Really, though, trait-eliciting conditions should be divided into objective and subjective components, as I argued above. Someone may be in a situation where helping is the appropriate response but not see it that way; conversely, someone may be in a situation where helping would be inappropriate but believe she ought to help.

As I pointed out above, both virtue ethicists and situationists agree on the importance of subjective states, and both also agree on the ultimate importance of behavior. Causally impotent virtues are not worthy of the name. As Adams (2006, p. 119) says, "surely a disposition to honest behavior is at least necessary, if not sufficient,

for a virtue of honesty." As long as virtue ethics maintains that there is a right thing (or a range of right things) to do for a given person in a given context, the precise details of the causal path from objective stimulus to behavior are in a way irrelevant. In the end, if one fails to do the right thing, one is not fully virtuous.37 In her defense of virtue ethics against the situationist challenge, Annas (2003, pp. 26-27) argues that a virtue is disposition to act and make choices, not just to behave; she stresses that "the agent's practical reasoning is essential." Yet she herself in articulating a theory of virtue (1993, p. 43) claims that the "virtuous person is," among other things, "the person who does in fact do the morally right thing." Presumably she is right that virtues are not mere behavioral dispositions, but that is because they are more, not less. Introducing the intervening variable of construal between objective stimulus and behavior just gives a fuller account of how people can fail to react virtuously; it does not save virtues from empirical critique. Thus, the ultimate objects of correlation remain objective conditions and behavior, but subjective construals form part of the theory connecting the two, much as genotype is part of the theory connecting the phenotype of parents with the phenotype of offspring. It is striking that many defenders of virtue ethics appeal to the power of construal as though it supported their defense against the situationist critique (Sreenivasan 2002, p. 58). Even the most recent monographs that explicitly address this debate (Russell 2009, Snow 2008) say so. As the foregoing discussion shows, however, the power of construal interferes with subjects' ability to respond consistently in a virtuous way to all virtue-eliciting circumstances.

---

[37] See Blum (1994), Driver (2001), McDowell (1979), and Rosati (1995).

6.2. Selection bias and arbitrary coherence

Though they argue that people are not cross-situationally consistent in the way that talk of traits leads us to believe, situationists usually also admit that, when socially embedded in day-to-day life, our attributions of traits lead to correct predictions of behavior. Like an unsound argument with a true conclusion, our reasoning processes begin with false premises about the existence and robustness of traits and derive true predictions about others' behavior. As Ross & Nisbett (1991, p. 7) put it, "biased processing of evidence plays an important role in perceptions of consistency," yet "the predictability of everyday life is, for the most part, real."[38] We use "fast and frugal heuristics" (Gigerenzer 2007, p. 158; see also Gigerenzer et al. 2000 and Sunstein 2005) to make inferences to conclusions that are (usually approximately) true in the environments we typically inhabit. Such heuristically powered inferences are not deductively valid arguments from true premises, and they can go wildly wrong if used in circumstances for which they were not adapted, but they do an adequate job of guiding us in everyday life.

Think of the old yarn about the King of Siam's refusal to believe that water could freeze: he made an inference from the behavior of water in typical conditions to the behavior of water in atypical (for him) conditions. We can reconstruct his reasoning as

---

[38] Doris (1998, p. 508) puts it even more strongly: "[S]ituationism is not embarrassed by the considerable behavioral regularity we do observe: because the preponderance of our life circumstances may involve a relatively structured range of situations, behavioral patterns are not, for the most part, haphazard."

relying on the premise that all water is liquid or vapor. While this premise is false, it led to only true inferences until the King contradicted his Dutch guest. Analogously, we make inferences based on traits about how people will behave in counterfactual scenarios (Edgar is honest; if anyone were honest he would not lie; so Edgar will not lie). Though the trait-invoking premises of such inferences are false, they will not lead us astray if the right social and other environmental influences conspire to make our conclusion true. We can safely fail to heed Marx's (1859/1904, p. 10) dictum that "it is not the consciousness of men that determines their social being, but on the contrary, their social being that determines their consciousness," provided we confound the two whenever we actually draw an inference.

Another way in which selection bias influences our perception of consistency, and thereby our estimation of the existence and robustness of traits, is through a phenomenon called arbitrary coherence. As Ross & Nisbett (1991, p. 19) put it, people "feel *obliged*, even committed, to act consistently." Ariely, Loewenstein, & Prelec (2003, p. 77; see Andrade & Ariely 2009, p. 5 and Ariely, Loewenstein, & Prelec 2006, p. 8) have argued persuasively that a person's "foundational choice" becomes "part of that person's stock of decisional precedents, ready to be invoked the next time a similar choice situation arises." They found that people's ordinal *rankings* of goods and experiences are generally stable (transitivity of preference is preserved) but that the *absolute level* at which the goods and experiences are rated is easily manipulated. For instance, students at the Harvard Business School were asked to write down a random two-digit number, then decide whether they would pay that amount for either of two

different bottles of red wine – one clearly superior to the other. After coming to a decision on this hypothetical question, the students actually bid on the bottles. All of them bid higher for the superior bottle (thus the coherence of their preferences), but the amount they bid was significantly affected by the two-digit number they had written: the higher the number, the more they were willing to pay (thus the arbitrariness).

Indeed, arbitrary coherence is not related merely to preserving transitivity of preference and (apparent) rationality of desire; it goes as deep as reasons to act: people who reject an offer as unfair in the ultimatum game are more likely to make fair offers when the tables are turned (Andrade & Ariely 2009, p. 6). This phenomenon may help to explain why situationist experiments have found stability (people do the same thing in the same circumstances) without consistency (people fail to do the same thing in different circumstances): we are able to navigate narrowly defined situations because of arbitrary coherence, but when we shift to a new situation we must first establish a new baseline; only then does arbitrary coherence kick in and create a new situational norm.

6.3. Availability bias and availability cascade

When people use the availability heuristic, they take the first few examples of a type that come to mind as emblematic of the whole dataset. It can lead to surprisingly accurate conclusions (Gigerenzer 2007, p. 28), but it can also lead to preposterously inaccurate guesses (Tversky & Kahnemann 1973, p. 241). We remember the one time Maria acted benevolently and forget all the times when she failed to show supererogatory kindness, leading us to infer that she must be a benevolent person.

In her defense of virtue ethics, Kupperman (2001, p. 243) mentions word-of-mouth testimony that "the one student who, when the Milgram experiment was performed at Princeton, walked out at the start was also the person who in Viet Nam blew the whistle on the My Lai massacre." Such tales are comforting: perhaps a few people really are compassionate in all kinds of circumstances, whether the battlefield or the lab. But while anecdotes about character may be soothing, it should be clear that anecdotal evidence is at best skewed and biased, as well as prone to misinterpretation. Perhaps a longitudinal study of this student would have found him compassionate only north north-west; when the wind was from the south he might not have known a moral hawk from a handsaw.

An availability cascade occurs when the availability bias goes viral; Kuran & Sunstein (1999, p. 683) define a cascade as "a self-reinforcing process of collective belief formation by which an expressed perception triggers a chain reaction that gives the perception of increasing plausibility through its rising availability in public discourse." In the echo chamber of Fox News, for instance, an insane claim about President Obama's wanting to institute "death panels" or being born in Kenya can be repeated so many times that it seems plausible. Especially when a news organization inspires people to repeat its talking points, which it then reports as news – a tactic for which Fox is well known – availability cascades can lead to massive misperception of the plausibility or probability of some claim. This phenomenon is relevant to our belief in traits because cascades about a person's personality are easily triggered. In fact, we already have a word for them: gossip.

6.4. The fundamental attribution error

According to Ross (1977, p. 183; see Ross & Nisbett 1991, Yzerbyt et al. 2001), people are prone to what he provocatively calls the "fundamental attribution error," a tendency to attribute most or all observed behavior to internal, dispositional factors rather than external, situational factors. When we observe others reading a script, for instance, we tend to assume they believe what they are saying, even when we are told in advance that they did not prepare the script and are merely reading it because they were asked to (Jones & Harris 1967, p. 22). We seemingly cannot help making what Uleman et al. (1996, p. 211) call "spontaneous trait inferences," which occur "when attending to another person's behavior produces a trait inference in the absence of our explicit intention to infer traits or form an impression of that person."

Why exactly humans have this innate, spontaneous reaction has not been fully answered, but its efficacy goes unchallenged. The problem may be partly perceptual, stemming from the Gestalt phenomenon: we focus on the figure rather than the ground. So when we observe other people acting in a situation, the people themselves are our focus and hence the only factor we consider in explaining their behavior (Harman 1999, p. 325; Ross & Nisbett 1991, p. 139).[39]

---

[39] The obverse side of the fundamental attribution error has to do with people's attributions regarding their own behavior. According to Jones & Nisbett (1971, p. 93), the unique breakdown of the fundamental attribution error occurs when we explain what we ourselves have done: instead of underemphasizing the influence of environmental factors, we overemphasize. Especially when the outcome is negative, we attribute our

6.5. The false consensus effect or egocentric attribution bias

Ross, Greene, & House (1977; see also Fields & Schuman 1976) were the first to discuss the false consensus effect, which occurs when people assume that their own opinion, desire, or other internal state is representative of the opinions, desires, etc. of a group or population. In particular, this effect engenders a tendency to think that one's own choice in a particular dilemma is the norm. For instance, Ross and colleagues asked passers-by on the street to carry a sign inscribed "Eat at Joe's!" Those who agreed to help estimated that 62% of others would do the same, while those who declined thought others would decline 67% of the time. Unless 29% of people would both accept and decline the solicitation, something is fishy about these numbers. The false consensus effect helps explain away intuitions about dispositions in the following way: once we make the fundamental attribution error and attribute a trait, we assume everyone else attributes the trait too, thereby reinforcing our own belief. If we assume that others explain Ignacio's tax-evasion as expressing his dishonesty, we are more likely to say so to them, thereby triggering an availability cascade about Ignacio, who might be an otherwise upstanding citizen. If, however, we agree with members of the

---

actions to external factors. This bias seems to tell against situationism, since it suggests that we *can* recognize the power of situations at least in some cases. However, the existence of the actor-observer bias has recently come in for trenchant criticism from Malle (2006), whose meta-analysis of three decades worth of data fails to demonstrate a consistent actor-observer asymmetry. See also Malle, Knobe, & Nelson (2007).

growing Tea Party movement in the United States that tax evasion is patriotic, we may praise him as a civic hero, triggering a different cascade.

6.6. The base rate fallacy

According to Bayes's Law, the conditional probability of a hypothesis given some evidence (the "posterior") is equal to the product of the prior probability of the hypothesis and the conditional probability of the evidence given the hypothesis (the "likelihood"), divided by the prior probability of the evidence:

(6.1) $$\mathrm{P}(h \mid e) = \frac{\mathrm{P}(h) \cdot \mathrm{P}(e \mid h)}{\mathrm{P}(e)}$$

The base rate fallacy occurs when someone neglects the prior probabilities, thereby over- or under-estimating posterior probability. People are notoriously bad at probabilistic reasoning, committing the base rate fallacy even after they have been trained to recognize and avoid it.

This fallacy is relevant to intuitions about dispositions because it leads us to jump from a small number of observed trait-relevant actions to the attribution of the full-fledged trait. Let *h* be the proposition *Karla is honest* and *e* be the proposition *Karla does not lie at time* t. Presumably $\mathrm{P}(e \mid h) \approx 1$, so if the prior probabilities of P(*h*) and P(*e*) are ignored (i.e., assumed to be 1), the posteriori P(*h* | *e*) is also roughly 1. Conditionalizing on a single example of truthfulness thus leads to an inflated estimation of Karla's honesty. If instead one recognized the validity of Diogenes' lantern, one would take into account the low prior probability of *e* (say .6) and even lower prior probability of *h* (say, .0001), leading to a less sanguine view of Karla's character:

$$(6.2)\quad \mathrm{P}(h \mid e) = \frac{\mathrm{P}(h) \cdot \mathrm{P}(e \mid h)}{\mathrm{P}(e)} = \frac{.0001 \cdot 1}{.6} = .00017$$

As it turns out, a little knowledge really is a dangerous thing.

6.7. Anchoring and disregard of regression to the mean

Anchoring is the bias of relying solely or too heavily on a single, early-acquired piece of information when making a decision. In a study mentioned above, Ariely, Loewenstein, & Prelec (2003) found that when people wrote down the last two digits of their social security number then made bids for goods such as wine and chocolate, those who had written higher numbers submitted bids that were 60% to 120% greater than those submitted by people who had written low numbers. Of course, they themselves did not believe that their social security number influenced their bids, but that just goes to show the deliverances of introspection should be taken with about three barrels of salt. Often, the first thing we notice about a person serves as our anchor, so if Lenny does something vicious when one first meets him, one will expect him to act viciously in the future, and if Manny does something virtuous when one first meets him, one will expect him to act virtuously in the future.

Regression to the mean is a statistical phenomenon: when a variable is observed at several standard deviations from the mean initially, it is quite likely to be observed closer to the mean on subsequent occasions. Disregarding regression to the mean, then, is the fallacy of assuming an extreme variable will remain extreme on later observations. It has been noticed, for example, that flight instructors praise pilots for successfully maneuvering and criticize them for unsuccessfully maneuvering.

Regression to the mean should lead one to expect that, regardless of the praise or criticism, pilots would perform worse after a successful maneuver and better after an unsuccessful one. Failure to recognize regression to the mean, however, may lead flight instructors to attribute improvement to their criticism and worse flying to their praise, then conclude that praise makes pilots lazy, whereas criticism helps them to fly better.[40]

Similarly, if we see someone act justly, recognizing regression to the mean should lead us to expect her to behave with less than perfect justice on the next occasion, but – because of anchoring and disregard for regression – we instead expect her to continue exemplifying the ideal of justice. And conversely, if we witness someone act selfishly, recognizing regression to the mean should lead us to expect him to behave less selfishly in the future, but – because of anchoring and disregard for regression – we instead expect him to continue expressing the trait of selfishness.

6.8. Confirmation bias

Confirmation bias is the tendency to search for, interpret, and remember information in a way that confirms one's beliefs. The bias has a long pedigree, having been identified (though not under its current name) by Sir Francis Bacon (1620, p. 79), who said:

---

[40] Kahnemann & Tversky (1973, p. 251) were the first to use this example, but they had only anecdotal evidence. More recently, Dorsey-Palmeteer & Smith (unpublished) have corroborated Kahnemann & Tversky's story with hard data from US Navy flight training.

> 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 or despises, or else by some distinction sets aside or rejects.

Darwin (1887/1958, p. 44) in his autobiography remarks that after years of working as a scientist, he adopted the following "golden rule": "whenever a published fact, a new observation or thought came across me, which was opposed to my general results, to make a memorandum of it without fail and at once; for I had found by experience that such facts and thoughts were far more apt to escape from the memory than favourable ones."[41]

More recently, contemporary psychology has provided experimental evidence for the confirmation bias and its relation to trait ascription. Mischel & Peake (1982, p. 750) suggest that anecdotal evidence is likely to be biased (or misconstrued) because instances of prototypic trait-relevant behavior are given too much weight in assessments. When once we decide (perhaps because of the fundamental attribution error) that someone is cowardly or temperate or conscientious, all our further observations are guided and colored by that decision.

---

[41] See Popper (2002, p. 45) on the related tendency of scientists to interpret all data in light of their current theory and therefore find confirmation everywhere (and disconfirmation nowhere).

6.9. Intuitions explained away

With so many biases stacked against us, it should come as no surprise that we think people possess character traits. The existence of these biases does not prove that no one has traits. What it instead shows it that *regardless of whether people have traits* folk intuitions would lead us to attribute traits to them. Those who subscribe to a sensitivity analysis of knowledge, according to which one knows that *p* only if one would cease to believe *p* if *p* were false, may therefore conclude that we cannot know (at least not without the aid of scientific psychology) whether we possess character traits. Even those who think sensitivity is a necessary but insufficient condition for knowledge should be dubious about whether we can know we have traits on the basis of folk intuitions.

Chapter 7. The defense of virtue

Keep your friends close, and your enemies closer.
~ Sun Tzu, *The Art of War*

The situationist attack on virtue ethics is not without opponents. Three primary responses can be identified:

(7.1) Virtue is a rare ideal, so empirical data showing that most people are not virtuous is irrelevant.

(7.2) Although the situationist critique shows that broad-based traits do not exist, an empirically informed theory of virtue can still be formulated in terms of virtuous actions or local traits.

(7.3) The empirical data do not support the situationist critique.

Call these broad arguments the *dodge*, the *retreat*, and the *counterattack*, respectively. In this section, I canvass versions of all three arguments and assess them in light of the ten items included in the hard core of virtue ethics. Most of these arguments turn out either to be unsound or to concede part of the hard core of virtue ethics. A few developments of the counterattack and retreat look promising, however, and raise the issues of the *portability of context* and *situation-consumerism vs. -producerism*. If situationists are right that context largely determines human behavior, then behavior can be controlled by resisting the influence of context, by arranging to find oneself in the appropriate context, or by actively producing appropriate contexts.

7.1. The dodge

Advocates of the dodge point out that for ancient philosophers like Plato and Aristotle, full virtue is a rare last fruit of a lifelong project. That most people are not virtuous is therefore no surprise; in fact, it may be a prediction of a suitably elitist virtue ethics. Burnyeat (1980), for example, discusses the myriad ways in which an acolyte to virtue can go wrong, strongly suggesting the rarity of virtue before the situationist challenge was widely discussed in philosophy.[42]

Athanassoulis (2000, p. 218) and Miller (2003, p. 379) flesh out this argument by drawing on the Aristotelian distinctions among virtue, continence, incontinence, and vice. The virtuous person does the right thing wholeheartedly. The continent person does the right thing, but only conflictedly. The vicious person does the wrong thing wholeheartedly. The incontinent person, like the continent, is conflicted, but his better self loses the battle and he does the wrong thing. Swanton (2003, p. 30) points out that participants in the Milgram (1974) obedience experiments, a favorite stalking horse of situationists, displayed intense emotional distress. This is evidence, she claims, of nascent or incomplete virtue, not of lack of virtue. Obedient subjects were incontinent; they wanted to do the right thing but failed. Under the assumption that most people are either continent or incontinent, the power of seemingly insignificant situational factors poses less of a threat to virtue ethics.

---

[42] Though, as Gilbert Harman has taken the pains to emphasize to me, related theories had been promulgated by Sartre, Goffman, and Arendt.

The dodge may seem successful, but there are reasons to worry. First, it appears to cede the egalitarianism, explanatory power, predictive power conjuncts in the hard core of virtue ethics. If moral education is as hard as Burnyeat (1980) makes it seem, most people could not be brought to behave in accordance with virtue, let alone be virtuous. If most people are non-virtuous, then, *pace* MacIntyre's (1984, p. 199) claim that much of human behavior would be "genuinely inexplicable" without appeal to virtues, the virtues are loose cogs in our motivational machinery, reliably licensing neither the explanation nor the prediction of behavior. Second, if most people are non-virtuous, then moral education may involve the very moral schizophrenia that virtue ethicists have criticized in other ethical theories. If moral education takes the form of advising someone to do what the virtuous person would do, learners would be forced to ask "Is what I propose to do what the virtuous person would do in my situation?" rather than "Is my maxim universalizable?" or "Will my action maximize happiness?"

7.2. The retreat

7.2.1. Virtuous acts

If one agrees that the dodge gives up too much, one might then be tempted to retreat to an act-virtue theory like Thomson's (1996, 1997). Hurka (2001, 2006) goes so far as to say that the act-virtue theory should not be considered a retreat from its dispositional cousin; instead, he argues, people really recognize only virtuous acts as intrinsically valuable, relegating virtuous dispositions (if there are any) to mere instrumental value. Moving virtue ethics in this direction strips it of most of its appeal,

however, since it gives up on the consistency, stability, explanatory power, and predictive power of virtues.

7.2.2. Local virtues

A less extreme retreat endorses Doris's (2002, p. 62) theory of local virtues. If virtues are natural kinds, then they should be discovered *a posteriori*. Though ethicists may be disappointed to find that the global trait of honesty does not exist, they should be excited to find that there are stable local traits. Parsimonious lists of cardinal virtues notwithstanding, if it turns out that there are hundreds of virtues, so be it. As Kamtekar (2004, p. 479) points out, Aristotle himself paved the way for this idea when he distinguished greatness of soul from due pride (1123b) and magnificence from generosity (1122a). If theorists of local traits are right, people should aim not so much to develop robust global traits like courage but to reinforce local traits like courage in-the-face-of-physical-danger-while-in-a-good-mood and see to it that they are only called upon to act courageously when they are in the appropriate circumstances.

The practical problem for the theory of local traits is that virtues are typically individuated by their characteristic reasons (Russell 2009). Generosity appeals to the needs of others; courtesy appeals to conventions of society; courage appeals to a threat against something valuable. Local traits are individuated by both their characteristic reasons and the causal powers of the situation. Generosity while-in-a-good-mood appeals to both the needs of others and one's subjective state; courtesy while-watched appeals to both conventions of society and the presence of bystanders; courage in-the-face-of-physical-danger appeals to both threats against something valuable and the

perceived type of the threat. If local traits really are as fine-grained and individualized as Doris suggests, then a virtue theory framed in terms of local traits would have to modify the consistency condition (since as currently formulated it appeals only to reasons) and possibly reject the access condition (since there will end up being so many traits that we could not possibly determine what they are). Perhaps both changes are acceptable, but they do come at a cost.

7.3. The counterattack

7.3.1. Introspection

The counterattack comprises a slew of independent ripostes. The first and most easily disposed of comes from Annas (2003, p. 23), who claims that the deliverances of introspection confirm the existence of character traits – objective measurements by social scientists be damned. It should be easy to see, however, that introspection cuts no ice here. The question is not whether people take themselves to have traits but whether they actually do have traits.

7.3.2. Equivocation

The next counterattack comes from Sreenivasan (2008, citing Webber 2006a). He begins by pointing out that that the meaning of the term 'cross-situationally consistent trait' is a function of the principle of situation individuation. An agent is consistent in all situations of type *S* if and only if she reacts in the same way whenever she is in *S*. In the previous section, I argued that the principle of individuation relevant to virtue ethics relies on the reasons there are for someone to act. If this is right, then

cross-situational consistency for virtue ethics amounts to responding in the same way to all situations that provide the same reasons. Oddly, Sreenivasan (2008, p. 604), claims that "a trait is cross-situationally consistent if it is manifested across situations that differ in respect of the kind of feature inviting behaviour that manifests that trait," contrasting this with a principle of individuation according to which a trait is cross-situationally consistent if it is manifested "across situations in which this feature remains constant, but other features vary." As pointed out above in section 2, neither of these principles of virtue-individuation is satisfactory. The appropriate principle of individuation is more coarse-grained: it appeals to reasons.

7.3.3. Morally unimportant behavior

Next, consider the counterattack according to which situationist experiments do not test morally crucial behavior. As Sabini & Silver (2005, p. 540) say, "picking up or not picking up" a stranger's papers is not "a very important manifestation of a moral trait." While this is a fair criticism of the Isen & Levin (1972) phone booth experiment, it ignores a large swath of the situationist literature. Consider just the Darley & Batson (1973), Milgram (1974), and Haney, Banks, & Zimbardo (1973) studies so often referenced by both situationists and defenders of virtue ethics. Is the failure to help a distressed man lying by the side of the road morally trivial? Is applying dangerous shocks not morally crucial? Is forcing guiltless fellow-participants in an experiment to clean latrines with their bare hands morally unimportant? While some experiments test for unimportant behaviors, many test important ones. Worse still, evidence from behavioral economics suggests that people's actions in low-stakes situations reliably

predict their actions in high-stakes situations (Camerer & Thaler 1995; Hoffman, McCabe, Shachat, & Smith 1994; and Roth, Prasnikar, Okuno-Fujiwara, & Zamir 1991). If these results can be extrapolated to non-economic behavior, then the studies that use trivial actions as independent variables do in fact support skepticism about traits.

7.3.4. One-off vs. longitudinal studies

Another counterattack claims that one-off experiments cannot provide evidence for or against virtue. Sreenivasan (2008, p. 607) advances such an argument, saying that the question of consistency cannot be answered by comparing different cohorts under different experimental treatments. Instead, the only evidence they would accept looks at the same cohort under different experimental treatments over the course of days, weeks, months, or even years. He categorically denies that "any data from a one time performance experiment [can] do anything to establish" the conclusion that the participants in the experiment lack a character trait. Such data are not thick on the ground in the psychology literature. However, what little there are (e.g., Hartshorne & May 1928) lend no comfort to this counterattack. Worse still, Sreenivasan's denial seems to fly in the face of well-regarded scientific methodology: separating a sample into comparable sub-samples, subjecting each sub-sample to a single treatment, and inferring that what sub-sample 1 does in treatment 1 at time *t* is what any sub-sample would do in treatment 1 at time *t`*. Consider the parallel denial that any data from a one-off performance experiment can do anything to establish the conclusion that a sample of oil is insoluble in oil. Presumably what one sample of oil does when mixed with water is

ceteris paribus what any sample of oil would do when mixed with water, regardless of time. If this criticism is on the mark, the longitudinal study argument is unsound.

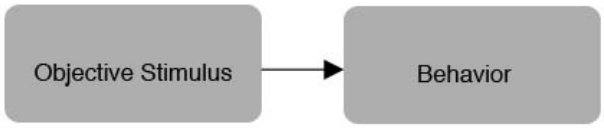
7.3.5. Confounding traits

The next counterattack argues that situationist experiments are confounded by further dispositional variables. They presume to test for one virtue, say honesty, but other virtues like prudence kick in and dampen the expression of the target trait. People with multiple, inconsistent character traits are bound to have opposing impulses and must violate at least some of them. Kamtekar (2004, p. 473), Miller (2003, p. 369), and Sreenivasan (2008, p. 607) claim that subjects in the Milgram experiments had two contrary virtues: compassion and obedience. Hence, when they did not express compassion it was not because they were the passive pawns of situational influences but because they were expressing the opposing virtue. Obedience may not be a virtue, but at least it is a trait. A less contentious version of this counterattack would therefore say that tests for trait t_1 may be confounded by the presence of traits $t_2, t_3, ..., t_n$, which may but need not be virtues. As an *a priori* exercise this argument is valid. What it needs, however, is empirical support. Milgram (1974) varied his experimental conditions to see whether people were simply obedient (as defenders of virtue now claim) or only obedient when certain situational cues were present. His results support the latter thesis. Subjects' obedience rate dropped to 0% when they were ordered around by the learner rather than the experimenter; they did not obey all orders, only the ones given by the experimenter, who was illuminated by the halo of authority.

7.3.6. The behaviorism bogeyman

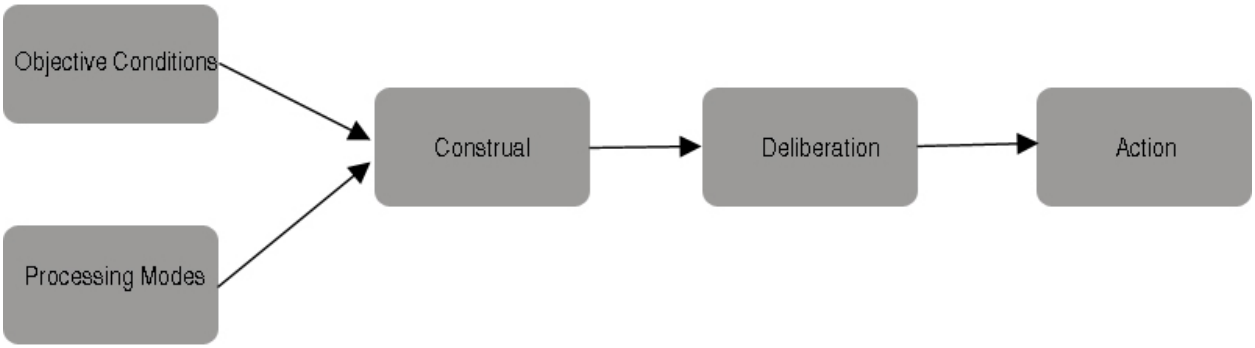
The penultimate counterattack argues that situationist experiments assume a crude – and refuted – behavioristic model of action. Correlations between objective situational variables and behavior fail to take into account subjective construal and deliberation. When this model is replaced by a nuanced, empirically adequate theory that refers to internal states and their interactions, it turns out that the data provided by situationist experiments can be either dismissed or explained away (Kristjansson 2008 and Upton 2009a, 2009b).

Figure 2: Stimulus-response arc model of action



However, situationists do not subscribe to the crude model of action mapped out in figure 2. Instead, they have a model like the one in figure 3, which includes at three other elements: construals, deliberation, and processing modes like moods.

Figure 3: Virtue ethical & situationist model of action



Processing modes influence attentional focusing; good moods, for example, induce attentional dilation and openness to new experiences[43] while bad moods induce the attentional constriction and closure to new experiences.[44] Without appeal to processing modes and subjective construals, situationism would lose much of its explanatory power, so it hardly makes sense to accuse situationists of behaviorism.

Furthermore, defenders of virtue ethics seem to misunderstand the situationist challenge when they locate subjective contruals outside the extension of 'situation'. Situationist psychologists and behavioral economists mostly seem to think of construals as part of one's situation. This may seem odd to those who think of 'situation' and 'external environment' as synonyms, but that is just because the experimenters in question have a more nuanced understanding of the *situation/non-situation* distinction. On the face of it, this distinction is straightforward: weather, presence of bystanders, social distance, ambient smells, ambient sounds, etc. are situational influences, whereas desires, values, and judgments are non-situational. The trouble with this way of understanding situational influences derives from the much broader extension granted to the term 'situational' by social psychologists and behavioral economists. For them, environmental properties are situational, but so too are properties like one's mood, emotions, and construal of the world. Though they do not give an explicit

---

[43] For more on the power of good moods, see Isen (1987) and Schaller & Cialdini (1990).

[44] For more on the power of bad moods, see Apsler (1975), Carlsmith & Gross (1968), Isen, Shalker, Clark, & Karp (1978), and Regan (1971).

definition of the distinction, these scientists regard seemingly internal, intentional features as situational. For the purposes of this section, I propose to work with a definition of the *situational/non-situational* distinction according to which a situational variable is (a) volatile and (b) etiologically connected with the agent's environment. Because our moods and emotions are so easily and quickly swayed by the weather, ambient smells, and ambient sounds, these too count as situational.

Construals are a trickier matter. Indeed, the term 'construal', though rampant in both the psychology and the philosophy literature, is ill-defined. In one sense of the term, a construal is just an agent's set of beliefs about the current state of the world. In another sense, a construal is an active choice to interpret ambiguous stimuli in one way rather than another. The latter type of construal obviously affects the former, and the former may affect the latter. In any case, the former seems to be what psychologists and behavioral economists have in mind when they call construals situational variables: immediate doxastic reactions to the environment, which are volatile and etiologically connected with the agent's environment. If this is right, the charge of behaviorism does not stick.

In any event, both virtue ethicists and situationists agree on the importance of both subjective states and, ultimately, behavior. Causally impotent virtues are not worthy of the name. As Adams (2006, p. 121) says, "surely a disposition to honest behavior is at least necessary, if not sufficient, for a virtue of honesty." The more complicated picture of action in figure 2 merely introduces new ways in which the connection between virte-eliciting conditions (which are individuated by reasons) and

behavior can be severed. In the end, if one fails to do the right thing, one is not fully virtuous. Annas, no friend of situationism, claims that the "virtuous person is," among other things, "the person who does in fact does the morally right thing" (1993, p. 43). One feels uncertain how this counterattack could ever have seemed plausible.

7.3.7. Parity of traits and situations

The final counterattack has two prongs. The first points out that the infamous 0.3 ceiling (Mischel 1968) on correlations between trait possession and trait expression is not a 0.0 ceiling, concluding that 0.3 might just be "sufficient to underwrite some conception of character" (Sabini & Silver 2005, p. 541). The second argues that situational influences have a correlation ceiling of their own, at roughly 0.4 (Funder & Ozer 1983), putting the power of dispositions and the power of situations on par. Bowers (1973) argued decades ago that the situationist critique overstates its case by treating the 0.3 correlation as negligible.

It seems that many toeing the situationist line have taken it for granted that if dispositions explain *N*% of behavior then situations explain (100 – *N*)%. This is emphatically not the case. What neither explains independently must be attributed to their interaction, to a third factor, or to randomness. Psychologists like Funder (2006) and Jost & Jost (2009) argue persuasively that the psychology of the 21st-century should supplement the recognition of the 0.3 ceiling with another that recognizes the limited power of situations. The new consensus weakens the situationist attack without entirely removing its sting. Combined with the second version of the retreat, however, it opens up the possibility of a virtue ethics couched in terms of the *portability of context* –

the degree to which agents may intentionally situate themselves in contexts that encourage (or at least do not discourage) behavior in accordance with virtue – and *situation-consumerism vs. -producerism* – taking an active role in shaping one's own situation and the situations of others, rather than viewing oneself as a passive pawn of situational influences.

7.4. The portability of context and situation-producerism

Suppose one has taken the situationist critique to heart, has given up on acquiring virtues individuated solely by reasons, and endorses instead a theory of local virtues in interaction with situations. What practical strategies might she employ? In this closing section I identify two strategies that can be used in tandem to coopt the power of situations.

7.4.1. The portability of context

One strategy involves the *portability of context*: seeking out situations conducive to one's particular situational susceptibilities. While Doris (2002, p. 147) is right that a romantic dinner with a flirtatious colleague while one's spouse is out of town is an easy sort of thing to avoid, what is to be said about ambient sounds, ambient smells, all-too-mercurial moods, and ever-changing social distance? Should jackhammers be banned? Should people always wear perfume? Should everyone take mood-enhancing drugs? Should people make sure always to look one another in the eye? The question is whether (or to what extent) a person can carry her preferred context with her.

Merritt (2000) attempts to answer the portability question with a theory of socially sustained virtues, arguing that the "sustaining social contribution to character" enables

people to act as if they possessed virtues. People behave as if they possessed global virtues not because they actually do possess them but because they find themselves in virtue-eliciting situation (a situation where reason *r* decisively calls for action in accordance with virtue V_r) only if they are also in the appropriate context (where situational forces conspire with local traits to bring one to act in accordance with V_r). She argues that, rather than trying in vain to galvanize your character against all temptations great and small, a more sensible [project] would be the exercise of care in your choice of [social situations]" (p. 378). The primary difficulty for Merritt's theory is that local virtues are not sustained solely through the social contribution; asocial situational factors like moods also affect virtue-relevant behavior. Thus, her theory should be considered an essential proper part of a full theory of local virtue.

Recall that according to globalist theories, for virtues to have explanatory power, subjunctive conditionals of the following form must be true:

(7.4) $V(a) \rightarrow \upsilon_a$

where *a* is an agent, *V* a virtue property, and υ_a the performance by *a* of an action in accordance with *V*. The theory of local traits strengthens the antecedent of the subjunctive conditional that connects virtue-possession with virtuous behavior. Let *C* be a context-property like *being in physical danger* or *being in social discomfort*, *a* an agent, V^C a *C*-relative virtue property like *courage in the face of physical danger* or *courage in the face of social discomfort*, and υ_a^C *a*'s acting in accordance with V^C. For the theory of local virtues to be correct, conditionals of the following form must be true:

(7.5) $[V^C(a) \wedge C(a)] \rightarrow \upsilon_a^C$

That is, if *a* possessed *C*-relative virtue *V* and *C* were true of *a*, then *a* would act in accordance with *V*. For example, suppose Wesley Autrey possessed courage in the face of physical danger (V^P) but not courage in the face of social discomfort (V^S). Were he confronted with a physically dangerous situation calling for courageous action, he would act appropriately:

(7.6) $[V^P(a) \wedge P(a)] \rightarrow \upsilon_a^P$

but were he confronted with a socially discomforting situation calling for courageous action, he would not:

(7.7) $[\neg V^S(a) \wedge S(a)] \rightarrow \neg \upsilon_a^S$

The trick, then, is to identify which local virtues one has (or could develop) and ensure that one remains in the appropriate contexts. One pragmatic use of research in social psychology, then, is the identification of such contexts. We may dream, for instance, that some day people will be able to take a virtue-battery, which will say which local virtues they have. With this knowledge, they could then plot out a life trajectory that (so far as possible) avoided situations incongruous with their trait signature. They could carry their preferred contexts with them – making use of the portability of context.

7.4.2. Situation-consumerism vs. situation-producerism

Another, more aggressive, strategy for dealing with the power of situational influences involves the distinction between *situation-consumerism and situation-producerism*. The portability response treats situations like restaurants. If Burger King is conducive to health, visit Burger King. If Taco Bell is not conducive to health, avoid

Taco Bell. If situation S_1 is conducive to virtue, seek situation S_1. If situation S_2 is not conducive to virtue, avoid situation S_2.

As I pointed out above, however, one's situation is not merely one's physical environment; it includes as well whatever is volatile and etiologically connected with one's environment. If we think of ourselves not only as situation-consumers but also situation-producers, the power of situational influences becomes a tool rather than a threat. To continue the health analogy, another way to approach nutrition is to make one's own meals. In the same way, rather than simply seeking and avoiding situations based on their virtue-conducive properties, we may take a more active role and create (both for ourselves and for others) situations with an eye to their virtue-conduciveness. One rather extreme example of this approach is Bentham's Panopticon, a building designed to make its inhabitants feel at all times as if they were watched. Another related example is ubiquitous use of surveillance cameras in European cities like London and Amsterdam. Though I have my reservations about these particular examples of moral technology, I do think they point the way to the kind of intervention consistent with the point of view of situation-producerism.

Marx (1845/1998) is famous for the final thesis on Feuerbach: "Philosophers have hitherto only interpreted the world in various ways; the point is to change it." Echoing that sentiment, we may conclude by saying that situationists and defenders of virtue ethics have hitherto only interpreted situations in various ways; the point is to change them.

Chapter 8. Fact, fiction, factition

Whatever they may think and say about their 'egoism', the great majority nonetheless do nothing for their ego their whole life long: what they do is done for the phantom of their ego which has formed itself in the heads of those around them and has been communicated to them; – as a consequence they all of them dwell in a fog of impersonal, semi-personal opinions, and arbitrary, as it were poetical evaluations, the one for ever in the head of someone else, and the head of this someone else again in the heads of others: a strange world of phantasms.
~ Friedrich Nietzsche, *Daybreak*, 105

Sir Walter [… had] been flattered into his very best and most polished behavior by Mr. Shepherd's assurances of his being known, by report, to the Admiral, as a model of good breeding
~ Jane Austen, *Persuasion,* chapter 5

'Cause I am, whatever you say I am
If I wasn't, then why would I say I am?
In the paper, the news everyday I am
I don't know it's just the way I am
~ Eminem, "The Way I Am"

I come not to bury virtue ethics, but to praise it.

In this section I argue that virtue (though not vice) attributions of the right sort should be made even if untrue. Drawing on formal work in multi-agent epistemic logic and empirical studies in social psychology, consumer research, and behavioral economics, I argue that the plausible, public attribution of virtuous traits induces both identification with those traits and belief that others expect one to act in trait-consonant ways, which in turn leads to trait-consonant behavior.

In section 8.1, I discuss the notions of placebo effects and self-fulfilling prophecies as instructive parallels to virtue-labeling. The beliefs involved in placebo effects and the announcements involved in self-fulfilling prophecies turn out to be true.

Nevertheless, an aura of perversity surrounds these beliefs and announcements, whose contents are true because they are believed, not believed because they are true. By inverting the direction-of-fit characteristic of belief and assertion, the beliefs and assertions involved in placebo effects and self-fulfilling prophecies arrive at truth through the back door.

Next, in section 8.2, I argue that thinking of virtue attributions merely as fact or fiction is too limited. We must recognize in addition a third category: factitious attributions, which make themselves true by being plausibly, publicly announced. Like the beliefs involved in placebo effects and self-fulfilling prophecies, those involved in labeling are true because believed, rather than believed because true. Along the way, I point out the difference between individual virtue-labeling and group virtue-labeling, arguing that there are game-theoretic reasons to expect the group variety to have a stronger effect.

In section 8.3, I address several potential objections – two descriptive, one normative – to the theory of virtue-labeling. The first is that research in moral licensing directly contradicts the theory of virtue labeling: praising people for their virtues actually leads to worse behavior, not better. A few studies seem to support this view, but the data on labeling and moral licensing can be fruitfully synthesized in a theory that recognizes the partial correctness of both views. The second objection is that the psychological construct of self-concept, which is implicated as a mechanism in the explanation of factitious virtuous behavior, is inconsistent with situationist insights. If this is right, then my argument is actually against situationism and for the existence of

character traits (albeit of a nontraditional sort). But, the objection goes, the weight of evidence so heavily favors situationism that there must be something wrong with my argument. I respond that self-concept alone is insufficient to simulate the virtues: the interpersonal forces of public announcement and mutual expectations are required to bring about factitious virtuous behavior. The first two objections are descriptive, arguing that there is no such thing as factitious virtue; the third is normative, arguing that factitious virtue is an immoral goal. According to this objection virtue-labeling essentially involves deception and therefore cannot be championed with a good conscience. I respond by arguing that virtue-labeling should be construed not as deceptive (knowingly false assertion) but as performative (disguised command).

8.1. Placebo effects and self-fulfilling prophecies

A primary norm of belief and assertion is truth. One should believe that *p* only if *p* is true. One should say that *p* only if *p* is true. On the face of it, there are only two ways for beliefs and assertions to relate to the truth norm. They can be *factual* (true) and hence satisfy it, or they can be *fictional* (false) and hence violate it. My contention is that the dimension of time (along, in many cases, with a social structure that enables announcements) introduces a third category: the *factitious*, which become true by being believed or publicly asserted.

8.1.1. Placebo effects

The placebo effect intrigues because the beliefs involved in it violate the truth norm initially yet satisfy it in the end. This phenomenon is often characterized as a belief about oneself causing its content to be true. Sarah believes at t_1 that her cancer

will go into remission, and at t_2 her cancer in fact does go into remission, due at least in part to her believing it would happen.

This is not the whole story of Sarah's recovery, however. The belief in question is usually arrived at through reasoning. She does not arbitrarily decide that she will be cured. A famous surgeon slices a chunk out of her. Or a cutting-edge radiologist zaps her with gamma rays. Or a televangelist prays over her. And on the basis of this intervention, she concludes that she will recover. We can rationally reconstruct the phenomenon by supposing that Sarah argues to herself as follows:

(8.1) If the televangelist prays over me, I will recover.

(8.2) The televangelist prays over me.

(8.3) Hence, I will recover.

This argument is valid, and – if the placebo effect occurs – it has a true conclusion. Yet we are justly skeptical of (8.1), especially if it is meant to have counterfactual validity.[45] The beliefs involved in placebo effects therefore satisfy the truth norm in a perverse way. They are not believed because they are true, but true because they are believed.

8.1.2. Self-fulfilling prophecies

Self-fulfilling prophecies, like the beliefs implicated in placebo effects, violate the truth norm initially but satisfy it in the end. They introduce a further element, however:

[45] The similarity to Gettier cases with false lemmas (Harman 1973) should be clear. When Smith infers that Jones owns a Ford or Brown is in Barcelona, he does so on the basis of a false belief about Jones. Similarly, when Sarah infers that she will get well, she does so on the basis of a false belief about the televangelist's powers.

public announcement. Recent work in formal epistemology has shown just how powerful such announcements can be in creating common knowledge (Chwe 2001) and common belief (Ditmarsch, Eijck, & Verbrugge 2009), determining which of several game-theoretic equilibria is played (Aumann & Brandenburger 1995), and generating further knowledge (Plaza 2007).

Most logics of public announcement deal only with truthful, or even knowledge-grounded, announcements.[46] A self-fulfilling prophecy, however, is neither of these. Were Fed Chairman Ben Bernanke to announce (arbitrarily, without any evidence) at a press conference on Sunday evening that the stock market would collapse the next day, people would react by selling their portfolios, leading indeed to a stock market crash.

The explanation of a self-fulfilling prophecy of this sort resembles the explanation of Sarah's mysterious recovery. Bernanke appears as a harbinger of doom on Sunday night, and people reason to themselves as follows:

(8.4) If Bernanke says there will be a crash tomorrow, then there will be a crash tomorrow.

(8.5) If there will be a crash tomorrow, I shall sell my portfolio right away.

(8.6) Bernanke says there will be a crash tomorrow.

(8.7) Hence, I shall sell my portfolio right away.

---

[46] Though see Van Rooy (2003) for a game-theoretic analysis of the power of communication in less-than-fully cooperative situations, where announcements cannot always be assumed to be true.

Then, when a large number of people act on the basis of the conclusion of this train of thought, they cause the market to crash.

The importance of the announcement to this sequence of events is paramount: only by seeing to it that everyone knew he expected the market to crash, and everyone knew that everyone knew that he expected the market to crash, and everyone knew that everyone knew that everyone knew that he expected the market to crash, etc. could Bernanke cause his announcement to turn out true in the end. Indeed, many or even all of the actors in this financial drama might think that Bernanke was lying through his teeth, but if most of them expect a sufficient number of the others to believe the announcement, then it still makes sense to sell their portfolios. They could reason to themselves as follows:

(8.8) If Bernanke says there will be a crash tomorrow, then everyone else will believe there will be a crash tomorrow.

(8.9) If everyone else believes there will be a crash tomorrow, they will sell their portfolios.

(8.10) If everyone else sells their portfolios, there will be a crash tomorrow.

(8.11) If there will be a crash tomorrow, I shall sell my portfolio right away.

(8.12) Bernanke says there will be a crash tomorrow.

(8.13) Hence, I shall sell my portfolio right away.

If Bernanke just silently thought to himself, “The market will crash tomorrow,” nothing would come of it. If he were to whisper it to just one person, nothing would come of it. Indeed, were he to whisper it to every stockholder individually (creating one level of

mutual knowledge but not common knowledge), a crash would be less likely. Only by making the announcement publicly would he generate the common expectations necessary to cause such panic.

This picture may still be overly simplified. A more nuanced reconstruction would assume there are five types of people: (A) those who take Bernanke at his word, (B) those who think there are too many people of type A, (C) those who think there are too many people of types A and B, (D) those who think there are too many people of types A, B, and C, and (E) those who do not believe Bernanke and do not think others will. (Of course, we can continue up the levels of pessimists indefinitely, but the point should be clear without adding further levels.) If there are enough people of types A, B, C, and D there will be a crash. Those of type A sell because they believe Bernanke. Those of type B sell because they believe (generally *in sensu composito*) that those of type A will sell (even if there are no people of type A). Those of type C sell because they believe (generally *in sensu composito*) that those of types A and B will sell (even if there are no people of types A or B). Those of type D sell because they believe (generally *in sensu composito*) that those of types A, B, and C will sell (even if there are no people of types A, B, or C). Those of type E are left with worthless portfolios.[47]

---

[47] See Keynes's famous discussion of the beauty contest in *General Theory of Employment, Interest, and Money* (2009, p.130). Compare with the following admission by Charles Prince III, the former CEO of Citigroup, in an interview with the *Financial Times* on July 9, 2007: "When the music stops, in terms of liquidity, things will be complicated. But as long as the music is playing, you've got to get up and dance."

In any event, the differences and similarities between the beliefs involved in placebo effects and the those involved in self-fulfilling prophecies bear emphasis. Whereas the initial (false) premise of the reasoning behind a placebo effect may be supplied tacitly, the corresponding premise in the reasoning behind a self-fulfilling prophecy cannot. Placebo effects essentially involve beliefs about oneself; self-fulfilling prophecies do not. Nevertheless, the perversity of self-fulfilling prophecies parallels that of the beliefs involved in placebos: they are true because they are announced, not announced because they are true.

8.2. Virtuous fact, fiction, factition

It is generally assumed that if virtue attributions are not fact, then they must be fiction, hence in violation of the truth norm of assertion. In this section, I argue that a third category must be recognized: factitious virtue attributions. Like the beliefs involved in placebo effects, factitious virtues depend on beliefs about oneself; like self-fulfilling prophecies, factitious attributions are not true prior to their announcement but become true by being publicly announced; like both, virtue-labeling inverts the ordinary direction-of-fit characteristic of belief and assertion.

8.2.1. Labeling and self-concept

In a seminal study, Miller, Brickman, & Bolen (1975; see Albarracín, & McNatt 2005, Burger & Caldwell 2003, Ouellette & Wood 1998, Strenta & DeJong 1981, Tybout & Yalch 1980, Vaidyanathan & Praveen 2005) compared the effects of labeling with

those of moral exhortation on the behavior of fifth graders.[48] Participants in the exhortation group were asked repeatedly by the principal, the teachers, and even the janitor to keep their classroom tidy. The labeling group, by contrast, heard a congratulatory (false) announcement of their above-average tidiness. After a brief increase in tidiness, the exhortation group settled back into its old routine, but the labeling group exhibited higher levels of tidiness over an extended period.

The favored explanation of this labeling effect appeals to the notion of self-concept: the set of beliefs one has about one's own personality traits. According to this explanation, the fifth graders who were labeled as tidy incorporated that claim into their self-concept, then acted accordingly. We can reconstruct their reasoning as follows:

(8.14) If the teacher says I am tidy, then I am tidy.

(8.15) If I am tidy, then I shall act tidily.

(8.16) The teacher says I am tidy.

(8.17) Hence, I shall act tidily.

The failure of exhortation to produce tidy behavior may also be explained in terms of labeling. There seems to be an implicature connecting present-tense normative statements and present-tense negative statements. Uttering statements of the form "*x* should be *F*," implicates statements of the form, "*x* is not-*F*," and conversely. Thus,

[48] To my knowledge, the only philosopher to cite this study in relation to the situationist challenge is Prinz (2009). His treatment is quite different from mine, as I discuss in section 4.2.

when the fifth graders were told that they should be tidy, it was implicated that they were not tidy, and they may have taken the negative label to heart.

As in the case of the self-fulfilling prophecy of financial panic, what was initially announced was not true until it was announced. In fact, in this case, the attribution of tidiness (where tidiness is understood as a counterfactual-supporting trait) never becomes true, but something very closely related does: that the students act tidily. Predictions on the basis of the trait, however, would be true; the students act tidily even though they are not tidy (in the sense of possessing a counterfactual-supporting, reasons-responsive trait). The false belief in the trait corresponds to Sarah's false belief in the televangelist's otherworldly power and shareholders' belief in Bernanke's predictive power; the tidy behavior corresponds to Sarah's recovery and the market crash. Perhaps it would be most accurate to say, then, that virtue-labeling causes the *factitious simulation* of virtues without causing factitious virtues.[49]

---

[49] Another way of construing the phenomena is to index the relevant proposition to times. Under this interpretation, the full content of a virtue-labeling announcement is "Agent *a* is virtuous at t_1," which is false. What turns out to be true is a proposition with the same content but a different time index: "Agent *a* is virtuous at t_2." However, trait-possession (and, a fortiori, virtue-possession) is not supposed to be the sort of thing that changes from moment to moment. An honest man does not become dishonest in the blink of an eye, and a coward does not become courageous overnight. Rather, if that appeared to happen, we would be inclined to say the person never possessed the trait to begin with. Thanks to Michael Levin for emphasizing this interpretive possibility.

In a number of other studies, the Miller, Brickman, and Bolen result was borne out. Jensen & Moore (1977), for instance, found that children labeled as charitable donated more than those subjected to moral suasion.[50] Grusec, Kuczynski, Simutis, & Rushton (1978) announced to experimental participants that a questionnaire they had completed indicated either that they were competitive or that they were cooperative, inducing congruent behavior in a subsequent game. Grusec & Redler (1980; see also Mills & Grusec 1989) found that ten-year-olds who helped once and were then labeled ("You know, you certainly are a nice person. I bet you're someone who is helpful whenever possible.") contributed 350% more in a subsequent trial than students whose actions were praised after helping ("You know, that was certainly a nice thing to do. It was good that you helped me with my work here today.").

At this point, an objection arises. One might be inclined to complain that these studies provide only weak support for the power of labeling because they were conducted with children. What reason do we have to think that labeling works on adults too? Consider first the fact that the strength of the labeling effect seems to be a function of age: older children are more susceptible to it (Grusec & Redler 1980). In addition, similar studies have succeeded in inducing labeling effects in adults. Allen (1982) showed that labeling a whole population in a television advertisement ("American consumers are willing participants in solving the energy problem.") increased their intention to conserve. More recently Cornelissen et al. (2006) showed that labeling

[50] Interestingly for those concerned with cross-situational consistency, this effect held in a variety of situations.

adults as eco-friendly proved more successful in inducing cooperative environmental activity than providing them with information about the environmental effects of their behavior. In another study, Cornelissen et al. (2007; see also Albarracín & McNatt 2005, Burger & Caldwell 2003, Ouellette & Wood 1998, Tybout & Yalch 1980, Vaidyanathan & Praveen 2005) found that labeling even led participants to reinterpret their past behavior as motivated by the trait, and that labeling was especially effective when the label was applied while the participants' cognitive resources were distracted. Presumably this is because they were unable to deliberate about whether to accept the attribution, so they took it at face value and uncritically incorporated it into their self-concept.[51]

Another potential objection is that my thesis is too strong: labeling does not always work, so factitious virtue is not guaranteed by calling someone virtuous. Here it becomes necessary to refine the claim about factitious virtue, which is only induced by *plausible, public* announcements to an audience that has a *correct conception* of the virtue in question.

---

[51] Though they uncritically incorporate the attribution, it may not be entirely irrational to do so. As I point out in the next section, labeling is most effective when it is plausible – especially when someone has performed a trait-consonant act immediately prior to the trait attribution. It seems that people cast about for an evidential basis for accepting the attribution, but that they do not bother excessively about the quality of the evidence.

8.2.2. The plausibility condition

Intuitively, the announcement must be plausible for the agent to take it seriously. Telling Ebenezer Scrooge that he is generous would provoke a scoff, not a donation. Calling Glenn Beck temperate would not end his tirades. Research in the social sciences bears out this intuition. Labeling has been shown to be most effective when it comes directly on the heels of trait-consonant behavior (Scott & Yalch 1980, Cornelissen et al. 2007). Kraut (1973), for instance, found that individuals who were labeled as generous immediately after donating were more likely to donate two weeks later to a different charity.[52] Furthermore, trait labeling is especially effective when the label is consistent with the target's initial self-concept; it seems to bolster the relevant portions of the self-concept (Cornelissen et al. 2007). Tybout & Yalch (1980) showed that people who viewed themselves as political were especially responsive to labeling related to their voting habits. Also along these lines, Henderlong & Lepper (2002) found that labeling works best when it is perceived as sincere and when it attributes performance to controllable causes.

8.2.3. The publicity condition

For quite different reasons, the announcement must be public. Tacitly labeling someone as courageous will have no effect on her behavior. But calling her courageous in front of a crowd could put into her the very mettle being attributed. We

[52] Note the similarity between this and the so-called foot-in-the-door technique (Freedman & Fraser 1966).

can give stronger reasons to endorse the publicity condition, though. First, publicly labeling someone prompts her to believe in the attribution, thus triggering a placebo-like effect through the mechanism of self-concept.

Second, publicly labeling someone causes the audience of the announcement to expect her to behave in accordance with the label. And, by creating common knowledge of this expectation, it leads her to know that they expect her to behave appropriately, to know that they know that she knows that they expect her to behave appropriately, to know that they know that she knows that they know that she knows that they expect her to behave appropriately, etc. Consider the notion of a Nash equilibrium, where mutual expectations among people lead to the fulfillment of those very expectations. Roughly, an equilibrium is a state of affairs where everyone expects *p* to happen on the basis of their and others' choices, and no one has an incentive to prevent *p* as long as everyone else lives up to expectations. The standard case of such an equilibrium is driving on the right (or left) side of the road. Everyone expects everyone else to drive on the right (left), and everyone prefers to drive on that side provided everyone else does. When Hitler publicly decreed that Czechoslovakians would henceforth drive on the right side of the road, he laid the grounds for an equilibrium that far outlasted the Third Reich's conquest. If a person is not unconditionally virtuous but prefers to act in accordance with a virtue provided everyone expects her to do so and acts accordingly, then publicly labeling her with that virtue will lead to the equilibrium where she does so.

Traditionally, the virtues have been viewed as "*corrective*, each standing at a point at which there is some temptation to be resisted or deficiency of motivation to be made good" (Foot 1997). Thus, getting people to expect (in the normative sense of 'expect') virtuous behavior should not take more than a plausible, public announcement. It follows that if someone is a conditional norm follower, making such an announcement attributing a virtue to him will lead him to act in accordance with the virtue.

8.2.4. The correct conception condition[53]

Finally, for virtue-labeling to have the desired effect, the labeled person must have a roughly correct conception of the virtue. Labeling as reasonable a person with a history of unreasonable behavior who believed in his own reasonability would not induce a change in his behavior. It might even reinforce his unreasonable behavior.

8.2.5. The inadvisability of vice-labeling

If the foregoing is correct, it should be evident why virtue-labeling but not vice-labeling is justified. Virtue-labeling causes someone to view herself as virtuous and believe that others expect her to act virtuously, which leads her to want to act in accordance with virtue. Vice-labeling, by contrast, causes someone to believe in her own viciousness and in others' expectation that she will act viciously, thereby leading her to want to act in accordance with vice. Unfortunately, direct studies of negative labeling have yet to be performed. An empirical prediction of my theory is that they

[53] Thanks to Brian Robinson for pointing out this third condition on successfully inducing factitious virtue.

would show that plausibly and publicly calling someone (and not just someone's action) vicious would induce further actions in accordance with vice. I draw comfort, however, from studies that show exhortation to be largely inert because, as I argued above, exhorting someone to be *F* implicates that he is not-*F*, and thus serves as a proxy for negative labeling. Further indirect support for this view of vice-labeling can be drawn from investigations of the power of stereotypes to influence the behavior of the targets of stereotypes. Walton & Spencer (2009), for instance, found that students belonging to groups stereotypically considered stupid performed worse on standardized tests when they were reminded of their group identity.

8.2.6. Interpersonal forces in labeling

The previous section argued that the explaining the traditional virtue-labeling effect in terms of self-concept is on the right track but incomplete because it fails to mention the interpersonal forces at work in the production of labeling effects. In the Miller, Brickman, & Bolen study mentioned above, not only did each child believe "I am tidy," but also each believed both, "Everyone else is tidy and will act tidily," and, "Everyone else thinks I am tidy and expects me to act tidily." Indeed, it was common knowledge that everyone believed everyone else was tidy and would behave tidily. The mutual nature of this social reinforcement will become important in section 8.3.2., where I discuss the compatibility of virtue-labeling with situationism. Presently, however, there are two points to be made. The first is that self-concept alone may not be sufficient to induce a labeling effect. Believing oneself virtuous is not as strong a spur to virtuous

behavior as both believing oneself virtuous and knowing that others believe one to be virtuous.

The second, more important, point is that labeling an individual and labeling a group are crucially different. Bicchieri (2006; see also Bicchieri & Xiao 2009, and Bicchieri & Chavez 2009) has argued persuasively that many people fit the description of what she calls a *conditional norm follower* – a person who prefers to comply with a norm provided a sufficient number of others comply, and that they expect him to comply in turn (even when it hurts him materially). In a recent paper, Fischbacher, Gächter, & Fehr (forthcoming) argue on the grounds of a novel experimental design that roughly half the population has such preferences. The theories of Nash equilibria and conditional norm following suggest that plausibly, publicly labeling a whole group will induce more trait-consonant behavior than plausibly, publicly labeling each individual in the group. This is because simultaneous group-labeling leads to common knowledge of the expectation for all to cooperate, and many people prefer to cooperate only on the condition that all (or a sufficient proportion) of the others cooperate and expect him to cooperate as well.

To see the difference between individual and group labeling, consider the two necessary conditions of being a conditional norm follower. In the individual case, plausible, public virtue-labeling generates one of the necessary expectations. If someone is plausibly, publicly attributed the trait of charitability, common belief arises: everyone expects him to act accordingly, everyone knows that everyone expects him to act accordingly, everyone knows that everyone knows that everyone expects him to act

accordingly, etc. However, he has no reason to expect others to act in accordance with the norm themselves. In the group case, by contrast, plausible, public labeling generates both of the necessary expectations: each member of the group expects the others to act charitably, and each believes that the others expect him to act charitably. Thus, group labeling should induce more conformity to norms than individual labeling, though individual labeling in turn induces more conformity than exhortation or non-intervention.

Locating the virtues internally, as counterfactual-supporting, reasons-responsive character traits, has proven to be untenable. The time has come to investigate the interpersonal dynamics that enable and encourage virtuous behavior, what Merritt (2000) calls the *sustaining social contribution to character*. One aspect of this contribution is, I submit, virtue-labeling and its sustaining and bolstering effects on self-concept. Labeling shares with placebos the causal power of an agent's beliefs about herself over her fate and with self-fulfilling prophecies the essentiality of plausible, public announcements.

8.3. Objections to virtue labeling

"All that is well and good," responds an empirically-minded philosopher, "but you have neglected three important things. First, research into moral licensing shows that praising people for their morality actually leads to worse behavior, not better. Second, you claim that self-concepts have causal power, but one of the key insights of situationism is that only situational variables explain and predict behavior. And third, even if you are right that virtue-labeling works, it should not be done. You are arguing

for a sort of noble lie: people do not actually have the virtues, but we should tell them they do so that they behave themselves. That is despicable." In this section I address these three concerns, arguing that virtue-labeling can be defended against all of them.

8.3.1. Factitious virtue versus moral licensing

A recent spate of articles seems to militate directly against the theory of virtue-labeling. Monin & Miller (2001), for instance, found that supplying people with what they call moral credentials led them to behave badly later. To understand why their results are consistent with virtue-labeling, we must delve into the details of their study. In their first experiment, they prompted participants to agree or disagree with obviously sexist statements like, "Most women are not really smart."[54] The participants were then presented a role-playing vignette, which asked them to decide whether a man or woman would be better suited to a sales job in the cement manufacturing industry. The moral licensing effect reared its ugly head when participants who had previously disagreed with the obviously sexist statements (thereby establishing their credentials as non-sexist) said that the job would be better suited to men than women.

Monin & Miller's second experiment was similar, but involved racism rather than sexism. Each participant was given a chance to make what could be construed as a

[54] In fact, this statement is not by itself sexist. If someone also believed that most men are not really smart, he could agree to without a hint of prejudice. Monin and Miller are justified, however, in characterizing the statement as sexist in an indirect way: it was presented to participants without a corresponding statement about men or people in general, so it implicated an invidious gender-based distinction.

non-racist choice, then was presented with a role-playing vignette. This time, the participant was to imagine that he was a police chief in a small rural area of the United States, that he knew his officers were racist, and that he did not want to provoke the officers. In light of these constraints, would he hire a black officer? As in the first experiment, participants who had been given a chance to make what could be construed as an unprejudiced choice (and thus had established, at least in their own minds, their moral credentials) were more likely to say they would hire a white.

In neither of these experiments is there a plausible, public attribution of virtue. Instead, participants were merely given a chance to make a choice that could be construed as moral. While this may have had an effect on their self-concept, it may not have influenced their beliefs about others' expectations of them. Plausible, public announcements do not merely influence self-concept; they ensure that everyone is on the same page by creating common knowledge. Another important difference between this study and labeling studies is that participants in these experiments were asked to role-play; the questions were, "What would you do if you were the manager of the cement manufacturing plant?" and "What would you do if you were the police chief?" They were asked to take on the perspective and even the motivations of another. By contrast, the labeling studies cited above measured decisions people made for themselves. Finally, given the stated goals attributed to their roles in these vignettes, participants who made the sexist or racist choice were arguably correct. After all, in the cement-manufacturing vignette, the primary hiring criterion was the ability to establish contacts with (presumably sexist) foremen and building contractors, and in the police

vignette, “you do not want to provoke any major unrest within the [explicitly racist] ranks.”

Other moral licensing studies differ from labeling studies in the same ways. In Khan & Dhar’s (2006) experiment, for example, no plausible, public announcements were made. Participants were asked to role-play rather than make actual choices. Finally, the choice they made was between “vicious” designer jeans and a “virtuous” vacuum cleaner; unsurprisingly, the undergraduate participants flocked to the jeans, implicating the importance of the correct conception condition. Similarly, in the study conducted by Sachdeva, Iliev, & Medin (2009), participants were allowed to make choices and update their self-concepts in light of the choices, but no plausible, public announcements were used.

Rather than showing that virtue-labeling is a chimera, moral licensing studies emphasize the boundary conditions of the power of self-concept and the importance of plausible, public announcements to ensuring trait-consonant behavior. Labeling works not just through beliefs about oneself but through beliefs about what others expect from one. Self-concept is mediated by the sustaining social contribution to character (Merritt 2000).

8.3.2. Self-concept and situationism: Friends or foes?

I have argued that the conjunction of self-concepts and common belief is reasons-responsive and counterfactual-supporting. These are the very properties denied of character traits by situationism. Is Prinz (2009) right that self-concepts

answer the situationist challenge? Can personality traits be rehabilitated in the self-concept paradigm? I fear that the answer to both of these questions is "No."

There are three reasons for this negative response, two having to do with self-concept in general, one with the virtues in particular. First, because our self-concepts are so easily swayed by plausible, public announcements, they are internal variables only in an attenuated sense. They require frequent social reinforcement and maintenance, unlike traditional Aristotelian virtues. Second, understanding the difference between moral licensing and virtue-labeling enables us to see why self-concept alone cannot simulate the virtues. Self-concept alone does not reliably induce factitious virtuous behavior; only self-concept working in tandem with the interpersonal forces that govern conditional norm-following do that.

Third, it is widely agreed that virtuous people do not generally reason in terms of their own virtues ("Since I am honest…"), but in terms of the reasons to which virtues respond ("Since it would be a lie to say…"). Perhaps the most emphatic statement of this view is in Williams (1985), which I quote at length:

> the virtue-term itself usually does not occur in the content of the [virtuous person's] deliberation. Someone who has a particular virtue does actions because they fall under certain descriptions and avoids others because they fall under other descriptions. That person is described in terms of the virtue, and so are his or her actions: thus he or she is a just or courageous person who does just or courageous things. But – and this is the point – it

> is rarely the case that the description that applies to the agent and to the action is the same as that in terms of which the agent chooses the action.

By contrast, self-concept-based factitious virtues essentially require appeal to one's own traits. A pert illustration of this point is the joke about the rabbi who fell asleep on his deathbed. His students sat around the bed, lauding his many virtues. Eventually, he regained consciousness but pretended to sleep, listening with pleasure to their praise. During a lull in the conversation, he opened his eyes and said, "And of my modesty you say nothing?"

8.3.3. Damning with feigned praise?

I began by saying that I would praise virtue ethics. Plausible, public virtue-labeling leads to factitious virtuous acts even though people do not possess the counterfactual-supporting, reasons-responsive traits presupposed by virtue ethics. If this is right, the resolution of the situationist challenge lies not in resisting it but in co-opting it.[55] One might think, however, that my praise is so faint as to be damning. Should we, as I have argued, feign praise, intentionally lying to one another in order to

---

[55] See Sarkissian (forthcoming), who argues that we should view situationism not only from the point of view of situation-consumers but also from the point of view of situation-producers. From the former perspective, situationism may seem pessimistic, showing that people do not have virtues. But from the latter perspective, situationism carries the glad tidings that we are able to produce virtue-inducing situations with only "minor tweaks" to our behavior.

trick ourselves into behaving well? Would not following this advice essentially involve the vice of dishonesty?

One's answer to this question should intuitively match one's answers to parallel questions about placebo effects and self-fulfilling prophecies. There is something distinctly perverse about such phenomena. They invert the mind-to-world and word-to-world direction-of-fit characteristic of belief and assertion. Yet the beliefs involved in placebo effects and the announcements involved in self-fulfilling prophecies turn out to be true. Is this just the advantage of theft over honest toil?

In the same way, virtue-labeling inverts the word-to-world direction-of-fit characteristic of announcements, but those announcements are validated when they turn out to be true. If we think of labeling as a covert imperative speech-act, the tang of perversity may dissipate. When a schoolteacher says, "Billy, I'll see you in my office after school," the surface form of her speech-act is an assertion. Clearly, though, the illocutionary force of her utterance is imperative – roughly equivalent to, "Billy, come to my office after school." In the same way, saying, "Lloyd Blankfein, you are so very generous," looks like an assertion, but the illocutionary force may instead be imperative – roughly equivalent to, "Lloyd Blankfein, you had better donate your bonus this year!" The truth norm applies to assertions but not to commands, so if virtue-labeling is actually a covert imperative, it does not violate the truth norm.[56]

---

[56] Compare this approach with the popular expressivist move of re-formulating seemingly assertoric moral statements like "X is good," as expressive statements like, "I like X."

Plausible, public virtue-labeling before an audience with a correct conception of virtue leads to factitious virtuous behavior through the mechanisms of self-concept and conditional norm-following. Since self-concept alone may lead to moral licensing, thinking of oneself as virtuous is not equivalent to being virtuous. The noble lie involved in virtue-labeling may be construed instead as a noble command. I conclude, then, with a twist on the Parmenidean principle: it is necessary to speak and to think what *ought to be*.

Chapter 9. Gyges in the Panopticon

Do everything as if Epicurus were watching you.
~ Seneca, *Letters to Lucilius*

I am a man; I consider nothing human alien to me.
~ Terence, *The Self-Tormenter*

We help our friends move apartments, but do nothing for victims of famine in the antipodes. We would give anything for our family and lovers, but we ignore the pleas of charity organizations. From one point of view, this is unremarkable, bordering on analytic: we care for those we care about and show no concern for those who do not concern us. From another point of view, however, it is striking: do we really think that people deserve help only if they are friends or family? Was Terence mistaken to think that because he was a man nothing human was alien to him? In our evolutionary prehistory, physical distance and social distance largely coincided. Had our ancestors desired to help people in distant climes, they would have been unable. From a practical perspective, then, caring only for those in one's physical and social sphere had the same effect as caring for everyone. The rise of telecommunication and cheap, high-speed transportation, however, moots physical distance, laying bare the power of social distance and forcing us to question the legitimacy of that power

But what exactly is this thing called social distance? How robust is the metaphor with physical distance? Can I take a social trip to reduce the distance between you and me? What sense can be given to the notions of social *area*, social *velocity*, and social *momentum*?

This section begins with a discussion of three examples of moral philosophers who focused on social distance. Plato addressed it in the myth of the Ring of Gyges; Epicurus placed great emphasis on it both in his writings and in the moral technology he deployed in the Garden; and Bentham's Panopticon relied on asymmetries of social distance to control the behavior of prison inmates. Next, I present a novel theory of social distance in terms of *interaction*, *group identity*, and *information*. This theory – especially its information aspect – draws support from recent work by experimental psychologists and economists. I conclude with a discussion of an experiment that I myself ran to test the power of the interaction and group identity aspects of social distance.

9.1. Factitious virtue and social distance in the history of philosophy

Social distance, though new to many contemporary philosophers, was historically an important aspect of moral thought. In this section, I address a few noteworthy historical examples of philosophical attention to the phenomenon.

9.1.1. The ring of Gyges

An early example of theorizing about social distance occurs in Plato's *Republic* (359c-360d), where Glaucon recounts the myth of the ring of Gyges, an artifact with the power to make its bearer invisible. After finding the ring, Gyges contrived to visit the royal palace. "When he arrived, he committed adultery with the king's wife and, along with her, set upon the king and killed him. And so he took over the rule." Glaucon proposes that everyone is like Gyges, and hence that inferences based on attributions

of virtues like honesty, justice, and fealty are invalid. What really predicts someone's behavior is his social distance from others:

> Now if there were two such rings, and the just man would put one on, and the unjust man the other, no one, as it would seem, would be so adamant as to stick by justice and bring himself to keep away from what belongs to others and not lay hold of it, although he had license to take what he wanted from the market without fear, and to go into houses and have intercourse with whomever he wanted, and to slay or release from bonds whomever he wanted, and to do other things as an equal to a god among humans. And in so doing, one would act no differently from the other, but both would go the same way. (360a-c)

We can precisify this argument in the following way. If someone possesses a character trait like honesty, she is disposed to behave in trait-relevant ways in both actual and counterfactual circumstances. However, no one – even the seemingly virtuous – *would* behave in virtue-relevant ways in both actual and counterfactual circumstances. Instead, immunity from the monitoring gaze of others (and the potential for sanction that such monitoring entails) would overpower whatever virtues someone might have, leading him to behave abominably.

If this is right, it suggests that from a third-person moral perspective, civil society should be guaranteed not by giving the population reasons to act but by manipulating social distance cues. In late 2009, the New York City Metropolitan Transit Authority began running announcements in the subway that said, "If you see an elderly, pregnant, or handicap person near you offer your seat. You'll be standing up for what's right.

Courtesy is contagious and it starts with you." Such an intervention epitomizes the reason-giving paradigm. If the moral of the myth of Gyges is correct, however, a more effective intervention would involve somehow decreasing the social distance between subway riders, thereby leading sitters to make room for those who most needed a seat.

From a first-person perspective, if the Gyges conjecture is right, it suggests that a project of self-improvement should focus on avoiding anonymity rather than cultivating virtuous thoughts and motives. If one were able to engineer situations such that one's social distance to morally relevant others decreased whenever other-regarding actions were appropriate, one would reliably behave in accordance with the other-regarding virtues.

9.1.2. The statue of Epicurus

Epicurus seems to have taken this message to heart. He developed a piece of moral technology unrivaled to this day. The theoretical basis of this moral technology is nicely summarized in Seneca's letters to Lucilius, where he reports Epicurus as having said: "We need to set our affections on some good man and keep him constantly before our eyes, so that we may live as if he were watching us and do everything as if he saw what we were doing" (11.8). By voluntarily decreasing their social distance from a revered symbol of authority, Epicurean acolytes put themselves in exactly the type of situation mentioned above. Rather than giving themselves reasons to act, they followed Epicurus' advice to artificially decrease their social distance from the watcher, thereby forcing themselves to behave as their better angels would have them behave.

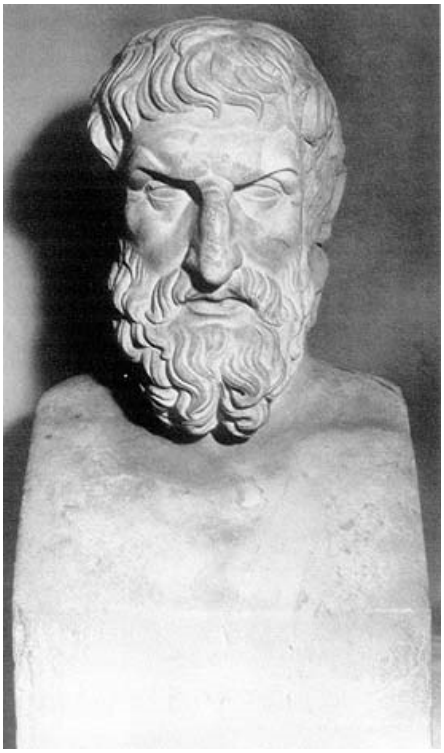
Interestingly, the moral technology behind this social distance intervention did not require that a real individual actually watch the Epicurean novices. Instead, they were to vividly imagine him as watching, which would trigger a privacy heuristics associated with real surveillance. At the end of this process of internalizing a revered monitoring authority, Epicureans needed no external props at all. In fragment 83, Epicurus is recorded as saying: "The man who has attained the natural end of the human race will be equally good, even though no one is present" (p. 51) and – presumably, watching. In psychoanalytic terminology: having introjected the revered watcher, they no longer need the fiction of an *other* who watches.

This theoretical basis became the groundwork for Epicurean moral technology. In light of the facts that a real watcher was unnecessary and that trainees needed to *feel* that they were watched, Epicurus used a surrogate watcher to guide his disciples on the path to virtue. Diogenes Laertius says in his *Lives of Eminent Philosophers* that many bronze statues of Epicurus were erected in Greece (1965, p. 537). The location and number of these bronzes is still a matter of contention, but almost certainly there was at least one in Athens. Furthermore, historians suspect that a statue of Epicurus stood in the Garden itself.[57]

---

[57] There is some contention as to whether the statue of Epicurus in the Garden was the *sole* Athenian statue of the philosopher. I tend towards Frischer's (1982) view that there were statues *both* in the Garden and in some public place (perhaps the Pompeion?), but my main hypothesis does not rest on this assumption.

Figure 4: Bust of Epicurus (Richter, vol. III, fig. 1153)

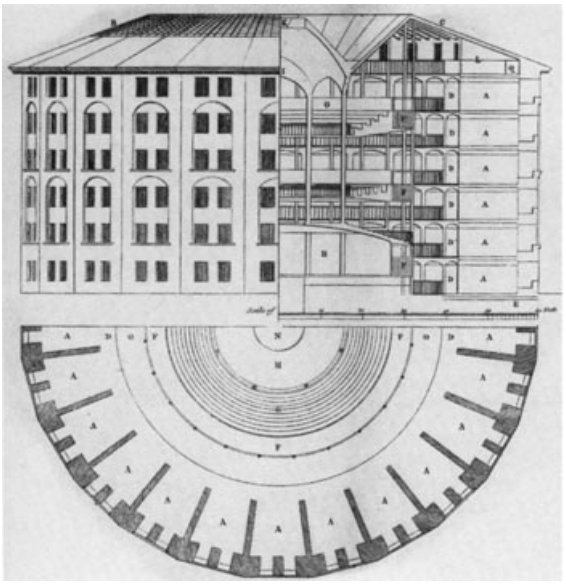


It is likely that this statue was commissioned by Epicurus, but at the very least it was commissioned by someone of his generation (Frischer 1982, p. 182). A statue, like a painting, tends to be experienced by those in its vicinity as a vigilant agency. And as we have already seen, the presence (real or fictive) of a monitor figured importantly in Epicurean moral training. It therefore seems highly likely that the statue of Epicurus in the Garden was set up for the express purpose of reminding his disciples of the sage's monitoring gaze. What better way could there be to ensure that his followers at all times kept in mind the maxim, "Do everything as if Epicurus were watching you" (25.5)? Constantly reminded of their master's fictive presence, and occasionally (during his life) reminded of his actual presence, the converts of Epicurus would constantly curb their actions as if someone were watching.

9.1.3. Bentham's Panopticon

My final historical example of moral philosophy's focus on social distance is Bentham's Panopticon, an architectural-geometric innovation. A building of this type consists of a hub of cells around an imposing guard tower. The cells have exactly two windows: one opening onto the outside of the ring, the other providing a vantage on the inside. In this way, the inmates are rendered incapable of communicating with their neighbors while the guard in the central tower is able to inspect many of them at once. What is more, the windows of the guard tower are shielded by Venetian blinds so that the prisoners cannot determine whether they are under the critical gaze of the guard.

Figure 5: Panopticon blueprint by Jeremy Bentham



It is a simple idea, but a brilliant one. By a few tricks of geometry and architecture, three asymmetries of social distance are guaranteed. First, the guard can see without being seen. Second, inmates are prevented from communicating with one another because radially the whole building lies open to view, but laterally the

perspective is constricted. Third, the guards' have full freedom of movement, while the prisoners' freedom of movement is severely hampered. In short: total ignorance and impotence from one point of view is paired with omniscience and omnipotence on the other. Bentham believed that situating the inmates of the Panopticon appropriately in the world of social distance would ensure that they behaved in a desirable way. He insisted not that their characters be reformed but that they be isolated and watched. "[T]he more constantly the persons to be inspected are under the eyes of the persons who should inspect them, the more perfectly will the purpose of the establishment have been attained. Ideal perfection, if that were the object, would require that each person should actually be in that predicament, during every instant of time. This being impossible, the next thing to be wished for is, that, at every instant, seeing reason to believe as such, and not being able to satisfy himself to the contrary, he should *conceive* himself to be so" (1995, p. 34).

Just as Epicurean acolytes internalized Epicurus as a monitoring agency with the help of the statue in the Garden, so the inmates of Bentham's Panopticon internalize the disciplinarian guards in the watchtower. As Foucault (1997, pp. 202-203) in his masterful study of the Panopticon puts it, "He who is subjected to a field of visibility, and who knows it, assumes responsibility for the constraints of power; he makes them play spontaneously upon himself."

While Plato emphasized the importance of social distance as such, Epicurus and Bentham invented moral technologies that used the power of social distance to regulate behavior. Epicurus' revered, symbolic watcher (the statue) and Bentham's anonymous,

unblinking watcher (the guard tower) stand at opposite poles of the surveillance spectrum, but their similarities are perhaps more important than their differences. By ensuring that their targets felt watched at all times, these interventions used social distance cues to regulate moral behavior.

9.2. Social distance today

Behavioral economists have recently begun struggling towards a definition of social distance. For instance, Hoffman, McCabe, & Smith (1996) define it as "the degree of reciprocity that people believe is inherent within a social interaction," and provide some evidence that social distance so construed is inversely correlated with giving behavior. This definition is clearly flawed, however. First of all, it builds a sort of infallibility into the definition of social distance. If the social distance between *a* and *b* is what *a* and *b* believe to be the distance between them, then of course they are right. But as anyone who has used a first name to address a superior when that superior clearly preferred to be addressed by his surname can attest, it is possible to underestimate one's social distance to another. Another reason to quibble with this definition of social distance is that even when people agree in their assessments of social distance, they may both be wrong. Consider two sworn enemies, both of whom have adopted clever disguises. They meet, fall in together, and think themselves good friends. Then each discovers the other's identity. "I was wrong about him all along!" each declares. They both perceived the social distance between them to be small, yet it was great. Finally, consider the fact that *a* may believe there is a high degree of reciprocity inherent in his relation to *b*, whereas *b* believes there is a low degree of

reciprocity inherent in their relation. They disagree. According to Hoffman, McCabe, & Smith, then, the social distance between them is undefined, since there is no one thing that is the degree of reciprocity they believe to be inherent within their interaction.

Bohnet & Frey (1999b, p. 44) proposed a revised definition, according to which "social distance is a much broader phenomenon that is not only relevant for social exchange-type relations but applies to all human interactions where some kind of other-regarding behavior is involved." This is an improvement but leaves much to be desired. Presumably social distance is a relation that holds all the time, not just when people are interacting in a potentially other-regarding situation. Just as there is a physical distance from Boston to Jakarta regardless of whether someone is planning to fly between them, so there is a social distance between people regardless of whether they are about to interact. In addition, Bohnet & Frey's definition provides no criteria for measuring social distance. For the metaphor of social distance to bear any weight, there must be criteria that enable us to answer questions of the form: Is *a* socially further from *b* than from *c*?

9.2.1. A theory of social distance

In this section, I articulate a definition of social distance in terms of *interaction*, *group identity*, and *information*. At first, I just dogmatically present the theory. Then I go into the details of each aspect of it, simultaneously explaining what they mean and what reason we have to include them in the definition of social distance.

The social distance *d* from *x* to *y* is a triadic relation (henceforth abbreviated $d(x, y) = z$), which is defined as a mapping from the Cartesian product of a domain *M* of agents to the real numbers:

(9.1) $d : M \times M \rightarrow \mathbb{R}$

such that $\forall x, y, z \in M$:

(9.2) $d(x,y) \geq 0$ [non-negativity]

(9.3) $d(x,y) = 0 \Leftrightarrow x = y$ [identity of indiscernibles]

The non-negativity condition simply ensures that social distance is like all other distance relations. Its justification is best understood via Groucho Marx's joke in *Duck Soup*: "If I were any closer, I'd be behind you." The identity of indiscernibles condition, which also parallels conditions for other distance relations, is a bit more speculative. It basically means that no one is closer to you than yourself. For those who experience a certain degree of self-alienation, however, this condition may seem dubious. Someone who sacrifices her own life to save a friend may be closer to that friend than she is to herself. Nevertheless, I will endorse the identity of indiscernibles condition here, reserving the right to revise it in light of further arguments and evidence.

In addition to the non-negativity and identity of indiscernibles conditions, the following qualitative *ceteris paribus* conditions govern *d*:

(9.4) if *x* can interact as easily with *y* as with *z*, $d(x,y) \leq d(x,z)$ [interaction-out]

(9.5) if *y* can interact as easily with *x* as can *z*, $d(x,y) \leq d(x,z)$ [interaction-in]

(9.6) if *x* shares at least as many group identities with *y* as with *z*, $d(x,y) \leq d(x,z)$ [group identity]

(9.7) if *x* knows or can know as much about *y* as about *z*, $d(x,y) \leq d(x,z)$ [information-out]

(9.8) if *y* knows or can know as much about *x* as *z* can, $d(x,y) \leq d(x,z)$ [information-in]

The reason for splitting the interaction and information conditions into their *out* and *in* directions is that, although being active, being passive, knowing, and being known all

have an effect on social distance, they have effects of different magnitudes (as I demonstrate with experimental evidence below). If you know about me but I do not know about you, we are both closer to each other than we otherwise would be, but I am even closer to you than you are to me. Unlike a classical distance metric, then, *d* violates the following two conditions:

(9.9) $d(x,y) = d(y,x)$ [symmetry]

(9.10) $d(x,z) \leq d(x,y) + d(y,z)$ [triangle inequality]

Social distance is asymmetric because *x* may be able to interact with or know about *y* when *y* is unable to interact with or know about *x*. For instance, in a jail the guards are able to initiate interactions with and can learn about the behavior of the prisoners much more easily than the prisoners can initiate interactions with or learn about the behavior of the guards. Social distance does not satisfy the triangle inequality because the friend of my friend need not be my friend at all. What follows is a motivating discussion of conditions (9.4) through (9.8) that draws on some recent studies in social psychology and behavioral economics.

9.2.2. Interaction and social distance

The ability to interact makes possible the application of sanctions, which may be rewarding or punitive, material or social. Reputation management becomes a pressing issue when others are able to retaliate or shun one for behavior they consider inappropriate or reward one for behavior they consider appropriate. Indeed, egoistic but rational utility-maximizers can be brought to see cooperation and following social norms

as their best option, provided they put some utility on the sanction they might receive for (not) cooperating or (not) following norms.

To investigate this idea, Kurzban, De Scioli, & O'Brien (2007) ran a study in which participants played a sequential prisoner's dilemma game (see table 5 below for typical payouts of a prisoner's dilemma game) in which, after each round, players were able to allocate some of their funds to a "punishment" account which would reduce the funds of the player with whom they had just interacted. Note that from a purely material point of view, using the punishment account is irrational. Since players interacted only once, they could not intimidate their partners into cooperating. Despite this fact, participants did in fact use their punishment accounts. Even more interesting, however, is the fact that fewer people defected when punishment was a live option than when it was not. The mere possibility of punishment made it unnecessary to punish. Perhaps most surprising, however, is the fact that players even engaged in third-party punishment. When given the opportunity to expend some of their own funds to decrease the funds of a player who had failed to cooperate with *someone else*, they chose to do so (though not as much as when they were given the chance to punish someone who had defected against them). Since players cooperated more when punishment was possible and even more when the punishment was coming from someone who might have a special reason to punish them, we have a prima facie case for the correctness of both interaction dimensions, and for the stronger effect of the *in* than *out* dimension.

Monetary punishment is not the only type of sanction; social sanction may also be leveraged to transform prima facie prisoner's dilemma games into coordination games.[58] Social sanctions are cheers and sneers, i.e., any form of non-material approval or disapproval. In addition, social sanctions are unlike market exchanges of other forms of utility in that they do not require two willing counterparties (the person being abused cannot simply decline it), are not fungible (you cannot buy and sell praise, or at least sincere praise), and cost very little to give. If social norms forbid defecting in a situation that looks from the material point of view like a prisoner's dilemma game, the introduction of social sanctions may transform it into a coordination game.

Table 5: Material payouts of prisoner's dilemma game

	Cooperate	Defect
Cooperate	$10, $10	$0, $20
Defect	$20, $0	$1, $1

In a game like the one illustrated in Table 5, the strongly dominant strategy for both players is to defect, despite the fact that when they employ this strategy they walk away with just $1 each. If we introduce the social sanctions of sneering or cheering, however, things change.

[58] See Bicchieri (2006), Brennan & Pettit (1993), Gächter & Fehr (1999), Lindbeck (1997), and Rege & Telle (2004).

Table 6: Material + social payouts of prisoner's dilemma game

	Cooperate	Defect
Cooperate	$10, $10	$0, $20
	cheer, cheer	cheer, sneer
Defect	$20, $0	$1, $1
	sneer, cheer	sneer, sneer

Provided the marginal utility of a cheer over a sneer is greater than the utility of $1, defecting ceases to be the dominant strategy in this game.

In a recent study, Masclet, Noussair, Tucker, & Villeval (2001; see also Rege & Telle 2004) ran a prisoner's dilemma experiment to test the power of both monetary and social sanctions. In the *no sanction* (*control*) condition participants could not sanction one another in any way. In the *monetary sanction* condition, participants could sanction one another by paying to decrease another's payoff. This condition therefore replicated the Kurzban, De Scioli, & O'Brien (2007) study mentioned above. Finally, in the *social sanction* condition, participants could sanction one another by assigning non-monetary punishment points. These punishment points did not affect players' material payoffs, but they did register disapproval. The question was whether simply knowing that they could be tagged with disapproval in this way would motivate participants to behave more cooperatively. In line with their predictions and with the theory of social sanction, Masclet, Noussair, Tucker, & Villeval found higher cooperation rates in both experimental conditions. In the monetary sanction condition, cooperation increased 85% compared to the control condition, whereas in the social sanction condition, cooperation increased 37% compared to the control condition. Though social sanction

did not induce as large an effect as monetary sanction, a 37% increase is far from negligible.

Interaction affects moral behavior beyond the wallet, as the Panopticon's prison bars illustrate. The proximity series in Milgram's (1974) studies in obedience also provides interesting evidence for the power of interaction. This experimental paradigm involves an experimenter, a confederate, and a subject. The experimenter tells the confederate and subject that they are participating in a study on the effects of punishment on learning. Through a rigged randomization mechanism, the confederate is assigned the role of *learner*, while the subject is assigned the role of *teacher*. The learner is strapped into a chair and fitted with electrodes, which are first tested on both the teacher and the learner to show that they give painful shocks. The teacher then quizzes the learner, whom he is prompted to shock after each wrong answer. Shocks start at 15 volts and increase by 15-volt increments.

> At 75 volts, the "learner" grunts. At 120 volts he complains verbally; at 150 volts he demands to be released from the experiment. His protests continue as the shocks escalate, growing increasingly vehement and emotional. At 285 volts his response can only be described as an agonized scream. […] At 300 volts the victim shouted in desperation that he would no longer provide answers to the memory test. […] At 315 volts, after a violent scream, the victim reaffirmed vehemently that he was no longer a participant. He provided no answers, but shrieked in agony

whenever a shock was administered. After 330 volts he was not heard from, nor did his answers reappear on the four-way signal box. (pp. 22-23)

If the teacher dissents, the experimenter replies politely but confidently, with an escalating sequence of prods. He first says, “Please continue,” or, “Please go on.” If the teacher dissents again, he is told, “The experiment requires that you continue.” If he dissents a third time, the experimenter says, “It is absolutely essential that you continue.” And if he dissents a fourth time, the reply is, “You have no other choice, you *must* go on.” If the teacher refuses once more after the fourth prod, the experiment ends. Otherwise, the experiment ends after the seemingly incapacitated learner is shocked three times at the maximum voltage of 450, which is labeled on the teacher’s dial merely as XXX.

The most astonishing result of the study is that a large majority of subjects were maximally obedient; that is, they failed to disobey five times consecutively and thereby end the study early. For our purposes, however, a secondary phenomenon proves quite interesting. Social distance was manipulated both between teacher and learner and between experimenter and teacher. Social distance between the teacher and learner was varied in the four treatments of the proximity series, during which social distance between experimenter and teacher was kept constant (and small, since the teacher was in the same room as the teacher):

“*Remote*. The learner is in a separate room and indicates his answers to the quiz questions by sending a signal of A, B, C, or D.

"*Voice-feedback*. The learner is in a separate room connected by an intercom.

"*Proximity*. The learner is in the same room as the teacher.

"*Touch proximity*. The learner is in the same room as the teacher; as the study progresses, the teacher must manually force the learner's arm onto the electrode in order to shock him." (p. 34)

As Milgram points out, the ability to interact increases from the remote condition through the touch proximity condition, in terms of both interaction-out and interaction-in. Predictably, as social distance increased, mean maximum shock level increased from 270 volts to 405 volts, a jump of 50%. While one may feel that even 270 volts as an appallingly large amount of electricity to put through another human being, it is much better than 405 volts.

Social distance between experimenter and teacher was varied in one further condition, where the experimenter left the room immediately after the experiment began and communicated only by phone. This served to increase social distance between experimenter and teacher, and it had profound effects: subjects reached a mean maximum shock level of 270 volts (the same as the touch proximity condition) and were fully obedient just 20% of the time (significantly lower than in the touch proximity condition, where full obedience was 30%).

9.2.3. Group identity and social distance

The Haney, Banks, & Zimbardo (1973) prison simulation supports the view that social distance as measured by group identity influences moral behavior. In this study,

participants were randomly assigned the role of either guard or prisoner. One would expect, then, that they would share most of their social identities. To heighten the sense that there was a difference between the guard identity and the prisoner identity, Zimbardo ordered different uniforms for each group. He reinforced the distinction between guards and prisoners by forbidding the use of prisoners' names; instead, each was addressed by the number sewn onto his uniform. By the end of the study, the prisoners had so internalized their new group identities that when they were offered "parole" (i.e., the chance to exit the experiment early) in exchange for forfeiture of their stipend, most accepted; then, when their parole application was "rejected," none left the experiment. They could of course have simply walked out, but they identified so strongly as prisoners that leaving without the consent of the experimenter was not a live option for them.

Zimbardo's experiment artificially increased social distance. By contrast, Epicurus' statue illustrates the use of perceptual cues to decrease social distance. Since everyone in the Garden was surveyed by the same statue, and the statue represented a figure they all knew and respected, group identity was affirmed and strengthened.

Data from behavioral economics also supports the group identity hypothesis. People are more generous with members of their in-group, even when it is common knowledge that the group was established by an arbitrary procedure immediately prior to the economic decision-making. Tajfel (1970; see Tajfel 1973, 1981, 1982), for example, established group identity by asking participants to estimate the number of

dots on a page. Those who overestimated were classed into one group, those who underestimated another. Members of the same group were significantly more generous towards and trusting of each other than were members of different groups. By focusing only on their most recently acquired group identity, they underestimated social distance from members of their own group and overestimated social distance from members of the other group.

Hoffman, McCabe, Shachat, & Smith (1994) varied Tajfel's paradigm in studies using the ultimatum and dictator games. They divided participants into groups based on their scores on a trivia quiz. Those who scored high and were given the chance to be dictators showed less generosity. In the ultimatum game, offerers who identified as high-scorers offered less to responders who identified as low-scorers, and the low-scorers accepted lower offers. Along these lines, Charness, Haruvy, & Sonsino (2007) tested the group identity hypothesis by having participants play economic games over the internet. They found that participants were less generous with one another when they did not share their national identities.

9.2.4. Information and social distance

Knowing about people and being able to find out more about them decreases social distance. Conversely, if others do or can know about one, social distance decreases as well. The limit case is one of free communication, where each person may learn anything she wants about the other and convey any information she wants to the other. In such situations, people may communicate social sanctions (sneers and cheers), strategy-relevant information, and personal information, each of which

influences moral behavior (Frey & Bohnet 1995). As mentioned above, Bentham's Panopticon paradigm institutes a fantasy of asymmetric omniscience where every twitch of a prisoner is observable by a guard. And as I discussed above, in the Milgram (1974) proximity series, when teachers were able to find out more information about learners (and, since this paradigm was symmetric, when learners were able to find out more information about teachers), destructive obedience decreased.

Along similar lines, Andreoni & Petrie (2004) found that providing participants in a public good game with information about the strategies of other players and photos of other players resulted in 59% higher contributions to the public good. Bohnet & Frey (1999a) used the dictator game paradigm to study giving behavior under four conditions:

(9.11) *Total anonymity*

(9.12) *One-way identification*: dictator knows who recipient is

(9.13) *One-way identification with information*: dictator knows who recipient is, as well as some information about her

(9.14) *Two-way identification*: both dictator and recipient know who the other is

They found that giving behavior and the informational aspect of social distance were inversely correlated: even though dictators in the one-way identification conditions were unidentifiable by recipients, in both of these conditions giving increased. The interaction aspect of social distance cannot account for this phenomenon, since it was common knowledge that recipients could not sanction their dictators. Merely knowing something about one's beneficiary (or victim) decreases social distance and encourages acts of generosity.

This phenomenon – when merely knowing the identity of one's potential victim or beneficiary increases pro-social behavior – was first discussed by Schelling (1968; see Eckel & Grossman 1996), who dubbed it the *identifiable victim effect*. Bohnet & Frey (1999b) conducted studies using both the prisoner's dilemma game and the dictator game, and found that mere silent identification of the beneficiary (but not the benefactor) was sufficient to induce fewer defections in prisoner's dilemma games and more altruism in dictator games. They explain this counter-intuitive result by saying, "One-way identification where potential benefactors receive some information on who their counterpart is induces solidarity by transforming an abstract, anonymous stranger into a visible, specified individual" (p. 53). Thus, their explanation of the information effect appeals to another dimension of social distance: group identity. Charness & Gneezy (2007) had participants play variants of the dictator and ultimatum games; they found that merely providing the surname of the receiver or responder increased the allocation made by the dictator or offerer.

These are just a few of the relevant studies. Others have shown that personal information decreases social distance more than impersonal information (Fox & Guyer 1978; Rege & Telle 2004) and that strategy-relevant information decreases distance more than strategy-irrelevant information (Hoffman, McCabe, Shachat, & Smith 1994; Rege & Telle 2004). In the Kurzban, De Scioli, & O'Brien (2007) experiment discussed above, it turned out that, "Knowledge that the experimenter, or the experimenter and other participants were going to know how much an individual punished increased this amount— more than tripling it in the latter case." All of these studies support Barclay's

(2004, p. 209) contention that "competition for scarce reputational benefits can help maintain cooperative behaviour because of competitive altruism" and thereby "solve" the tragedy of the commons, essentially transforming a mixed-motive prisoner's dilemma game into a coordination game.

Just as our assessment of shared group identities may be biased by recently acquired identities, so may our assessment of levels of privacy be influenced by unconscious indices of privacy. In particular, we systematically underestimate social distance when faces and face-like objects are in our field of vision (as Epicurus seems to have understood), leading us to act as if someone were watching us even when we are alone. Evolutionary arguments have been made to explain why humans are so keenly aware of faces and face-like objects in their environments (Barclay 2004); indeed, the fusiform face area of the brain seems to be a devoted module for facial recognition. Burnham & Hare (2007, p. 90) argue that "some of the anonymity effect in existing public goods experiments may be caused by activation of the dedicated neural architecture to detect faces. In other words, individuals may alter their level of prosocial behavior in the form of public goods contributions, in part, because of involuntary neural activation caused by the presence of human eyes and faces." Burnham (2003, p. 141) points out that one-way identification is evolutionarily anomalous, so it would have made sense for our ancestors to infer from "I see him" to "He sees me."[59]

---

[59] He goes on to mention that non-human animals also behave differently in the presence of faces and face-like objects: "Woodcutters in the forests of the Sundarbans in India are preyed upon by tigers. Noting that tigers always attack people from behind,

The Fox & Guyer (1978) study cited above used visual identification of one's partner to increase cooperation in a prisoner's dilemma game. In another fascinating study, Burnham (2003) varied the perception of privacy in a dictator game with three conditions:

(9.15) *Anonymity*. Neither participant knows about or sees the other.

(9.16) *Recipient photo*. The dictator privately sees a photo of the recipient.

(9.17) *Dictator photo*. The recipient privately sees a photo of the dictator.

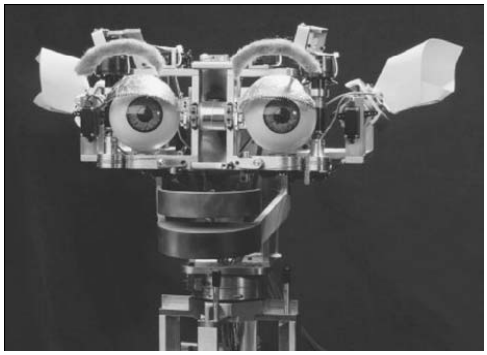
In the dictator photo condition, the interaction and information dimensions of social distance confound the perception of privacy: dissatisfied recipients know their dictators by sight, and may be able to find and sanction them outside the lab. In the recipient photo condition, however, neither of these confounds exists. Giving was enhanced in both experimental conditions, supporting both the *in* and *out* directions of the informational aspect of social distance hypothesis.

Burnham's results have been corroborated by a string of subsequent studies. Burnham & Hare (2007) used a photo of KISMET the MIT robot to trigger the facial recognition module of participants' brains.

---

a study was conducted where some woodcutters wore masks of human faces on the back of their heads. The study was halted in 1987 after 30 men without masks were eaten by tigers, versus zero in the control group. Note, that the study was halted because all men began wearing self-made masks (and still do)."

Figure 6: KISMET



Even though it was clear that they were not actually being watched, participants *felt* they were being watched. And KISMET had the intended effect, inducing an increase in dictators' allocations of about one third. Haley & Fessler (2005) replicated the KISMET experiment with a computerized dictator game in which the presence of "eye spots" on the computer's background induced extra giving. Bateson, Nettle, & Roberts (2006) replicated these results with a real-world experiment: they set up an honesty box in an academic tearoom to test whether perceived level of privacy would influence whether people paid in full for their beverages. On alternate weeks the experimenters put up either an unobtrusive photograph of flowers or a small photograph of human eyes. "On average, people paid 2.76 times as much [per cup] in the weeks with eyes." Presence of eyes explained 63.8% of the variance. Perhaps the most astonishing evidence for the perception hypothesis is Rigdon, Ishii, Watabe, & Kitayama's (2009) study, which used a minimal stimulus known to activate the fusiform face area of the brain.

Figure 7: Minimal facial stimulus

Merely presenting dictators with this arrangement of dots induced more giving, but when the figure was inverted, the effect disappeared. An exit interview with the dictators in this experiment found that they did not realize that they felt watched, even though their behavior showed clear signs of it. It seems that people can be brought to behave more generously through unobtrusive, unconsciously-processed stimuli.[60]

9.3. Short-circuiting social distance heuristics: An experiment

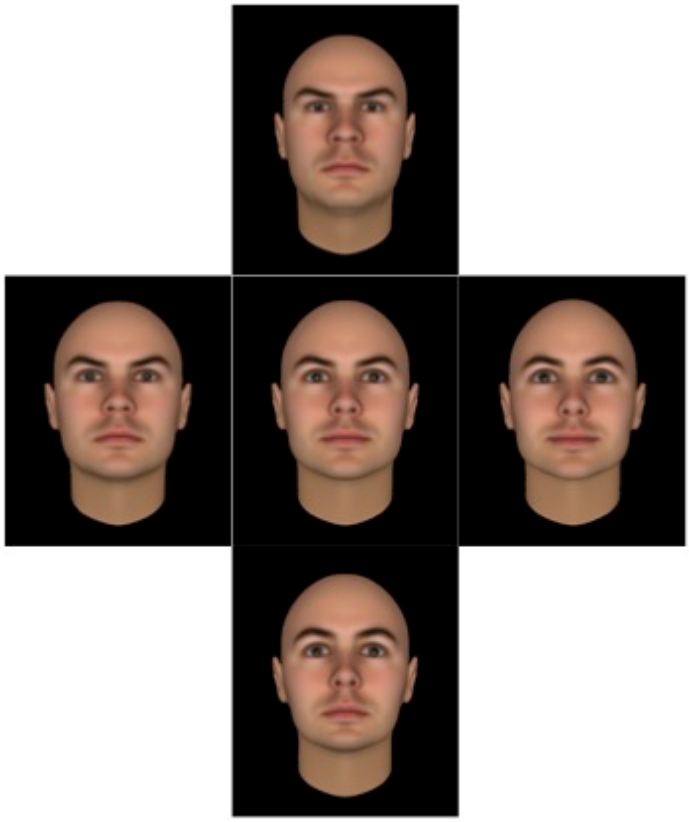
Since people can be brought to underestimate or overestimate social distance when its information dimension is short-circuited by KISMET, I decided to test the relative power of the interaction and group identity dimensions of social distance using a public good game. The experimental design crucially depends on Oosterhof & Todorov's (2008) study of the evaluation of faces on social dimensions. They found that evaluations of faces on social dimensions can be explained in terms of just two features: the dominance / submissiveness of the face, and the valence (trustworthiness /

[60] See Simmel (1964), who claims that eye contact has the sociological task of forming a bond between individuals.

untrustworthiness) of the face. For instance, a confident face is both dominant and trustworthy; a threatening face is dominant and untrustworthy.

Figure 8: Socially-valenced faces

(*y*-axis is dominance / submissiveness, *x*-axis is trustworthiness / untrustworthiness)



Oosterhof & Todorov have shown that the dominance / submissiveness axis is highly correlated with attributions of maturity / immaturity and masculinity / femininity. The more dominant the face, the more mature and masculine it looks; the more submissive the face, the more immature and feminine it appears. Thus, the dominance / submissiveness of a face corresponds to the interaction dimension: dominant faces appear more capable of interacting than submissive ones.

A parallel argument covers the valence of faces. Trustworthy (positive valence) faces make their viewers perceive enhanced group identity, while untrustworthy (negative valence) faces make their viewers perceive decreased group identity.

If this is right, then we should expect people who see a dominant or trustworthy face to underestimate social distance and people who see a submissive or untrustworthy face to overestimate social distance, in the same way that they overestimate social distance when KISMET watches and underestimate it when he does not.

9.3.1. The experimental design

To test this hypothesis, I conducted a public goods game ($n = 96$) with volunteer participants from an urban East-coast university. In this experimental paradigm, players receive 10 tokens per round over five rounds. They allocate the tokens to one of two envelopes – one marked “For me,” (the *Me* allocation) the other “For the group” (the *Group* allocation). Each token allocated to *Me* pays me $0.10 at the end of the game. Each token allocated to *Group* pays each player $0.05 at the end of the game, for a total of $0.30 combined. Thus, the most efficient strategy (the one that maximizes the overall payout) is to allocate all ten tokens each round to *Group*, but the dominant strategy (the one that maximizes the payout to oneself) is to allocate all ten tokens each round to *Me*.

Table 7: Payouts of public goods game in cents

	Others		
Self		Keep	Share
	Keep	10, $10 \times 5 = 50$	35, $25 \times 5 = 125$
	Share	5, $15 \times 5 = 75$	30, $30 \times 5 = 150$

Participants played in groups of 6, with each player assigned to a different condition:

(9.18) *Trustworthy*: tokens had a picture of a trustworthy face (see figure 5)

(9.19) *Dominant*: tokens had a picture of a dominant face

(9.20) *Neutral*: tokens had a picture of a neutral face

(9.21) *Submissive*: tokens had a picture of a submissive face

(9.22) *Untrustworthy*: tokens had a picture of an untrustworthy face

(9.23) *Control*: tokens had a picture of flowers (see figure 6)

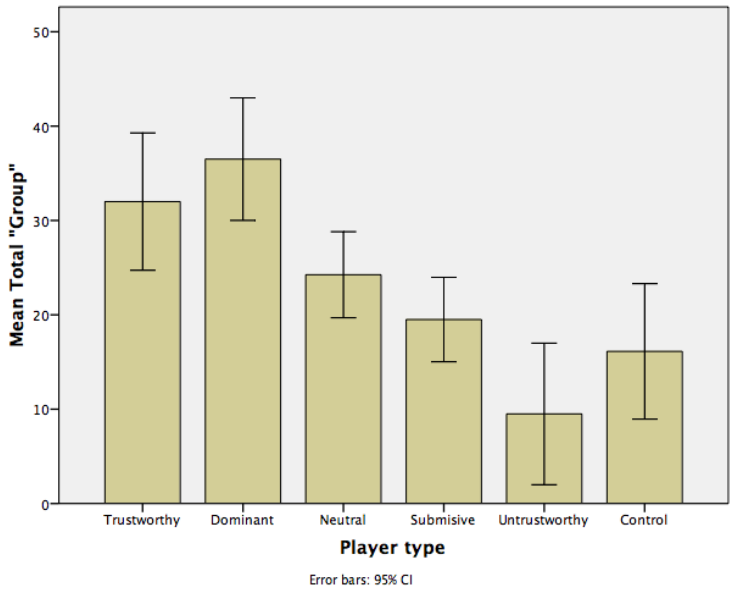
Figure 9: Control image (flowers)

The dependent variable was the number of tokens contributed to *Group*. If a player put all ten tokens in the *Me* envelope every round, this number would be 0 (payout to self = $5.00, payout to others = $0.00); if she put all ten tokens in the *Group* envelope every round, it would be 50 (payout to self = $2.50, payout to others = $12.50).

9.3.2. Results

Analysis of variance (ANOVA) revealed significant differences ($p < .01$) among the treatments. A post-hoc analysis showed significant differences between the trustworthy and dominant treatments on the one hand and all other treatments on the other hand. Participants in the trustworthy and dominant conditions contributed on average 73% and 64% ($p < .05$) of their tokens to the public good, compared to just 32% in the control condition. In addition, participants in the untrustworthy condition contributed just 19% of their tokens to the public good; while this low contribution was statistically indistinguishable from the contribution of the control condition, it did differ significantly from the neutral condition ($p < .05$). These results are represented graphically in figure 10.

Figure 10: Average contributions to *Group* (minimum 0, maximum 50)



In addition to the statistical significance of these differences, the effect sizes bear emphasis. Participants in the trustworthy condition gave roughly 100% more than those

in control, 60% more than those in submissive, and 200% more than those in untrustworthy ($p < .05$). Participants in the dominant condition gave roughly 125% more than those in control, 50% more than those in neutral, 80% more than those in submissive, and 260% more than those in untrustworthy ($p < .053$). Finally, participants in the neutral condition gave roughly 150% more than those in the untrustworthy condition ($p < .05$). These effect sizes are quite high compared to most of the effects canvassed above, suggesting that Oosterhof & Todorov really have hit a nerve in their analysis of faces on social dimensions.

9.3.3. Discussion

Whereas experiments like those of Bateson, Nettle, & Roberts (2006), Burnham & Hare (2007), Haley & Fessler (2005), and Rigdon, Ishii, Watabe, & Kitayama's (2009) focused on the information dimensions of social distance, my experiment used the interaction and group identity dimensions. To maximize giving behavior, we need to focus not only on *whether* people feel watched (as Bentham seems to have supposed) but on *what sort of person* they understand themselves to be watched by (as Epicurus seems to have understood). Since participants gave more in the trustworthy and dominant conditions and less in the untrustworthy condition, it appears that all three dimensions of social distance can be short-circuited using facial cues.

Further research should investigate the interactions of dimensions. For instance, how do people react to confident (dominant & trustworthy), threatening (dominant & untrustworthy), unconfident (submissive & untrustworthy), and unthreatening (submissive & trustworthy) faces? Do my results hold also in other types of games,

e.g., the sequential prisoner's dilemma? Is it an effective negotiating tactic to display a picture of a dominant face behind me when I talk with my counterparties?

We live in a network of social relations, often measuring our distance from others in degrees a familiarity rather than miles. Those degrees of familiarity, however, are highly volatile and prone to misinterpretation. My social distance from another depends on whether he sees me and whether I see him. My perceived social distance from another also depends on whether I feel watched – regardless of whether anyone is watching. The Gyges conjecture has been experimentally corroborated thousands of years after it was first put forward.

Not only that, but the wisdom of Epicurus' moral technology compared to Bentham's is now evident. While the Panopticon manipulates actual social distance, the statue in the Garden manipulates perceived social distance by short-circuiting privacy heuristics. Since the drawbacks of the Panopticon paradigm are staggering, future work in social control and self-improvement should follow Epicurus' lead by searching out new ways to alter people's perceptions of social distance without intervening in heavy-handed ways in their lives.

Chapter 10. Conclusion

Film is the ultimate pervert art. It doesn't give you what you desire; it tells you what to desire.
~ Slavoj Žižek, "The Pervert's Guide to Cinema"

[In a picture,] even the ignorant see what they ought to follow; in it the illiterate read.
~ Pope Gregory I, Letter 11:10

10.1. The pervasiveness of the situationist challenge:

Consequentialism in the crosshairs

The situationist challenge to virtue ethics has caused much ink to be spilled – a fair amount of it by me. Doris (1998, 2002) and Harman (1999, 2001, 2003, 2006) have argued that virtue ethics makes false presuppositions about the robustness of character traits like courage, courtesy, and honesty. If the challenge is successful, many assume, virtue ethics will have to be given up in favor of deontology, consequentialism, an ethics of care, or some other normative theory. Such an attitude reflects what I call the *containment thesis*: the situationist challenge is a crisis only for virtue theory. But perhaps the situationist challenge is like the sub-prime mortgage crisis. Perhaps we think it is contained while in fact it threatens to infect virtue theory's neighbors. If the containment thesis is false, consequentialists, deontologists, and other non-virtue theorists would do well to stop gloating at the straitened circumstances of virtue ethics and start worrying about their own vulnerability. In fact, I think precisely this is the case. In this section, I argue for the extension of the situationist challenge to

consequentialism, leaving its further extension to other normative theories for future work.[61]

10.1.1. Consequentialist theories of goodness, betterness, and rightness

Consequentialist theories are framed in terms of the consequences of action. An action φ is good if it has good consequences, better than some other action ψ if its consequences are better than the consequences of ψ, and optimal (and therefore right) if there is no action σ whose consequences are better than the consequences of φ. While there are of course many variations on this theme – some appealing to rules for action rather than actions as such, some appealing to direct or expected consequences rather than all or actual consequences, some appealing to goodness just for the actor, others to goodness for every sentient being, still others to something in between – the distinctions among these variants are irrelevant to my argument.

What makes consequentialist theories subject to the threat of situationism is their account of goodness and betterness. According to these theories, what makes the

---

[61] The capsule argument from situationism against deontology is that situational influences make the application of universalizable rules unreliable, invalidating the *ought implies can* commitment of most such theories. The capsule argument from situationism against the ethics of care is that the moral mindfulness and moral sensitivity advocated by such an ethics are traits like any other, so they stand or fall with the traits of character emphasized by virtue ethics.

consequences of an action good is that they satisfy preferences (so-called *preference utilitarianism*[62]) or cause pleasure and prevent pain (so-called *hedonic utilitarianism*[63]).

There are of course other views of what constitutes goodness. Some pluralist theories countenance preference-satisfaction, pleasure, and pain. Some countenance further elements of goodness beyond preference-satisfaction, pleasure, and pain, such as fairness, desert, beauty, knowledge, welfare, or capabilities.[64] My arguments in this section cover any consequentialism that defines goodness, betterness, and rightness at least in part by reference to preference-satisfaction, pleasure, or pain. It may be possible to extend the argument to cover other consequentialist theories, but I leave that question open.

With the target of my argument thus restricted, the problem can be stated succinctly: a recent insight of behavioral economics is that preferences, pleasures, and pains are indeterminate and subject to situational influences. These two features of conative states generate two arguments against consequentialism. The first is that

---

[62] Noteworthy preference utilitarians include Singer (1993) and Brandt (1979).

[63] Noteworthy hedonic utilitarians include Bentham (1789/1961), Mill (1861/1998), Feldman (1997, 2004), and Tannsjo (1998).

[64] Broome (1991, pp. 192-200) argues for the goodness of fair distributions. Feldman (1997, pp. 154-174) argues for the goodness of giving people what they deserve. Moore (1903, pp. 83-85; 1912) argues for the goodness of beauty and knowledge. Sen (1979) argues for the goodness of individuals' welfare. And Sen (1985) and Nussbaum (2000) argue for the goodness of capabilities.

consequentialist theories leave goodness, betterness, and rightness undefined, rendering them impotent. The second is that consequentialists have an untenable commitment to dynamic goodness, betterness, and rightness.

10.1.2. The arguments from indeterminacy and dynamics

Suppose goodness, betterness, and rightness are indeed indeterminate. Using the preference utilitarian model so that *being good* means satisfying preferences, and preferences are indeterminate, then what is good is indeterminate. If *being better than* means *satisfying more preferences than*, and preferences are indeterminate, then what is better it is indeterminate. If *being right* means sa*tisfying at least as many preferences as*, and preferences are indeterminate, then what is right it is indeterminate.

Turning to the hedonic utilitarianism, if *being good* means causing pleasure (no pain), and what would cause pleasure (pain) is indeterminate, then, again, what is good is indeterminate. If *being better than* means causing more pleasure than (less pain than), and what would cause pleasure (pain) is indeterminate, then being better than is indeterminate. And if *being right* means causing at least as much pleasure as (no more pain than), and what would cause pleasure (pain) is indeterminate, then rightness is indeterminate.

Hence, if behavioral economists are right that what is preferred and what causes pleasure and pain are indeterminate, then preference and hedonic utilitarianism are committed to the indeterminacy of goodness, betterness, and rightness. Presumably, though, any acceptable moral theory must be *complete* in the sense that any action one might reasonably consider is obligatory (right in a strong sense), permissible (right in a

weak sense), or forbidden (wrong) according to the theory. Both preference and hedonic utilitarianism, however, are incomplete: because preferences, pleasure-causation, and pain-causation are indeterminate, rightness according to these theories is indeterminate.

This argument can be seen as the metaphysical counterpart to an old epistemic warhorse often trotted out against consequentialism. According to the epistemic incompleteness argument, a moral theory must be *epistemically complete* in the sense that it must be possible to *know* of any action whether it is obligatory, permissible, or forbidden according to the theory. Since the full consequences of any particular action are unknowable, those consequences cannot be used to determine right action. The canonical response to the epistemic argument is to deny the necessity of epistemic completeness and argue that consequentialism is not a decision procedure but a standard of right and wrong.[65] Just as someone throwing a ball need not know the laws of physics for those laws to determine what happens to the ball, so a moral agent need not know the laws of morality for those laws to determine whether his action is right.

This response is unavailable as a defense against the metaphysical incompleteness argument. The point of my argument is not that we cannot know what the deliverances of consequentialism are but that it has no deliverances to offer in many cases. Of course, one could deny the completeness criterion anyway, but such a move is sure to raise eyebrows. Is it really true that an adequate moral theory may leave it

---

[65] Bentham (1789/1961, p. 31), Mill (1861/1998, 26), and Sidgwick (1907, p. 413) all deal with the epistemic incompleteness argument in this way.

open whether an action is obligatory, permissible, or forbidden? Imagine someone in a moral quandary with perfect information: he must choose between actions φ and ψ. A committed consequentialist, he consults his table of values to see whether the consequences of φ would be at least as good as the consequences of ψ. He knows what those consequences are, for all relevant agents, till the end of time. Yet when he attempts to weigh them against each other, the scales fall to pieces in his hands. Action φ is neither better than, worse than, or morally equivalent to ψ. Our poor agent is left undecided. This is the cost of denying completeness. Consequentialists who refuse to bite this bullet may prefer to challenge the evidence for the indeterminacy of preferences, dispositions to cause pleasure, and dispositions to cause pain, a strategy I consider below.

Next, consider the dynamics of goodness, betterness, and rightness. In addition to the completeness condition identified above, it would seem that any adequate moral theory must abide by the principle of *evaluative stasis*: if action φ is good (better / right) at time *t*, then φ is good (better / right) at time *t`*. There is of course an ambiguity in this formulation, leading to at least two readings. On the one hand, it could mean:

(10.1) If performing action φ at time *t* would be good at *t*, then performing φ at *t`* would be good at *t`*.

(10.2) If performing action φ at time *t* would be good at *t*, then performing φ at *t* would be good at *t`*.

Clearly, (10.1) is false. (10.2), however, is very plausible. The counterparts of evaluative stasis in terms of betterness and rightness are:

(10.3) If at time *t* performing action φ at *t* is better than performing action ψ at *t*, then at *t`* performing φ at *t* is better than performing ψ at *t*.

(10.4) If performing action φ at time *t* is right at *t*, then performing φ at *t* is right at *t`*.

Behavioral economists have discovered, however, that what is preferred (would cause pleasure) is a dynamic category. Our experiences change what we prefer and what we find pleasurable (painful) in later cases. Thus, according to both preference and hedonic utilitarianism, the moral status of an action at a time may change, violating the evaluative stasis principle.

This point bears emphasis, for it entails that if you do the right thing now, a revolution in preferences or pleasure-causation may transform what you did into the wrong thing, and if you do the wrong thing now, a revolution in preferences or pleasure-causation may transform what you did into the right thing. To clarify, this is not the simple point that people are often *mistaken* about whether they are doing the right thing. It means that they may actually do the right thing, which then later becomes the wrong thing to have done, and conversely. To counter this argument, consequentialists would have to deny the principle of evaluative stasis or debunk the economics research. I take it that the latter strategy is initially more appealing, so I now turn to the evidence for indeterminacy and dynamics.

10.1.3. Evidence for indeterminacy and dynamics

So far, I have merely asserted that behavioral economics has shown preference, pleasure, and pain to be indeterminate and dynamic. In this section, I explain the methodology and data behind this claim.

Let $V_a^t(p_s)$ be the value an agent *a* assigns at time *t* to proposition *p*'s obtaining at time *s*, which we can take to be one of the real numbers *n*.[66] Then *a*'s preference at *t* with respect to *p*'s truth at *s* is determinate just in case:

(10.5) $(\exists n)(V_a^t(p_s) = n)$

That is, *a* assigns at *t* some value or other to *p* at *s*. We may then say that *a*'s preferences are fully determinate at *t* just in case:

(10.6) $(\forall s)(\forall p)(\exists n)(V_a^t(p_s) = n)$

Generalizing on the agent and the time of evaluation, we may say that preferences as such are fully determinate just in case:

(10.7) $(\forall a)(\forall t)(\forall s)(\forall p)(\exists n)(V_a^t(p_s) = n)$

That is, everyone at all times assigns some value to every proposition's truth at any time. Any complete consequentialist theory that includes preference-satisfaction in its definition of goodness, betterness, and rightness is committed to (10.7).

---

[66] This way of construing preference presupposes that preferences are *atomic*, in the sense that the value assigned to one is independent of the value assigned to any other. Arguably, however, preferences are *holistic* in the sense that the value of one depends on the truth or falsity of others. In holistic preference models, the primitive bearers of value are not individual propositions but worlds. Fortunately, my argument here generalizes to holistic preferences simply by replacing references to propositions with references to worlds.

Research in behavioral economics, however, has provided evidence against (10.7). In one particularly striking study, Ariely, Loewenstein, & Prelec (2006) had participants write down the last two digits of their social security numbers (SSNs), thus generating an arbitrary two-digit number for each participant. Ariely and his colleagues announced that they would be holding an auction, then showcased a number of consumer items, such as computer peripherals, books, wine, and chocolate. Participants were asked to say for each item whether they would be willing to pay their SSN-truncation in dollars for the item. Hence, if someone had an SSN ending in 19, he had to say whether he would be willing to buy the wine for $19, whether he would be willing to buy the chocolate for $19, and so on. Another participant with an SSN ending in 83 had to say whether she would be willing to buy the wine for $83, whether she would be willing to buy the chocolate for $83, and so on. Next, participants were asked to bid on these same items, disregarding their SSN-truncations.

Since this was a real auction involving real money and real goods, the participants presumably put down numbers that reflected their actual preferences. Somewhat astonishingly, then, those with high SSN-truncations bid quite a bit more than those with low SSN-truncations – 57-107% more (p. 76). Ariely, Loewenstein, & Prelec take this result to show that participants did not assign values to the items until they were asked whether they would bid their SSN-truncations for the items. In other words, the participants violated (10.7) by not assigning a value to the items prior to the auction. When they were prompted to say whether they would pay their SSN-truncations for the

items, they formed preferences anchored around those SSN-truncations; then, when they were asked to bid, their bids reflected this anchoring effect.[67]

In another experiment, Ariely, Loewenstein, & Prelec showed that whether an experience is considered *preferable* (in preference utilitarian language, *good*) is subject to the same kind of anchoring effect. Participants were undergraduates in Ariely's marketing class at Berkeley. They were told that at the end of the semester Ariely would give a 15-minute reading from Walt Whitman's *Leaves of Grass*. Next, half of the participants (the *accept group*) were asked whether they would attend the reading in exchange for $2; the other half (the *pay group*) were asked whether they would pay $2 to attend the reading. Afterwards, all participants were given an opportunity to attend the reading for free. Only 8% of accept group signed up, compared with 35% of the pay group. This experiment suggests that preferences are indeterminate in an even stronger sense: people do not even assign a positive or negative valence to some potential objects of preference until prompted to do so.[68]

Preferential completeness was fairly easy to define. Hedonistic completeness, however, is a bit harder. Pleasure and pain are arguably orthogonal, in that the same

---

[67] This so-called *preference reversal* phenomenon has been replicated by Fischer & Hawkins (1993); Green, Jacowitz, Kahneman, & McFadden (1998); Johnson & Schkade (1989); Kahneman & Knetsch (1993); and Lichtenstein & Slovic (1971, 1973). See Slovic (1995) and Ariely & Norton (2008) for manifestos of preference-constructivism.

[68] To make matters worse, Ariely & Norton (2008) have shown that people do not realize that their preferences are indeterminate in this way.

experience may cause both of them, only one of them, or neither. However, many hedonic utilitarians put pleasure and pain on a single axis, counting pain as negative pleasure and pleasure as negative pain. This simplifying assumption is dubious, but I make it here because it plays no important role in the argument for hedonic indeterminacy. Let us say that the hedonic value at time *t* of experiencing *p* at time *s* for agent *a* is given by a function similar to the one for preference: $H_a^t(p_s)$. A positive value for *H* corresponds to pleasure, a negative value corresponds to pain, and a value of zero corresponds to the absence of pleasure and pain. Higher positive values correspond to more pleasure; lower negative values correspond to more pain. Given such an *H*, hedonic value is complete if and only if:

(10.8) $(\forall a)(\forall t)(\forall s)(\forall p)(\exists n)(H_a^t(p_s) = n)$

That is, everyone at all times assigns some hedonic value to experiencing every state of affairs at any time. Any complete consequentialist theory that includes pleasure and pain in its definition of goodness, betterness, and rightness is committed to (10.8).

Demonstrating the indeterminacy of hedonic value is a bit harder than demonstrating the indeterminacy of preference, but experimental economists have found some suggestive evidence nevertheless. The basic idea behind their methodology is to show that normatively irrelevant situational features influence the degree to which someone finds experiences pleasurable (painful). From this they infer that, prior to the experience, the agent did not assign a determinate hedonic value to the experience. In yet another experiment associated with their (2006), Ariely, Loewenstein, & Prelec began with the same SSN-truncation treatment. Participants

wrote down the last two digits of their SSNs. They then listened through headphone to an unpleasant noise: “a high-pitched scream (a triangular wave with frequency 3,000 Hz), similar to the broadcast warning signal” (p. 80). Next, they were asked to say whether they would be willing to listen to the same sound again for a number of cents equal to their SSN-truncations. After answering this hypothetical question, participants were given an opportunity to earn some money by listening to the unpleasant sound. They bid in a reverse auction on the obligation to listen to the sound: the higher the bid, the more they would be paid to listen. Once again, the SSN-truncation led to different bids. Those who had considered being paid a higher amount for listening made higher bids; those who had considered a lower payment made lower bids. This result suggests that the hedonic value of listening to the sound was indeterminate *even after they had listened to it once*. Only after they considered the hypothetical question about their SSN-truncations did participants form a determinate hedonic valuation of the experience.

In addition to the evidence for indeterminacy of preferences, pleasures, and pains, economists have conducted studies that suggest that preferences, as well as dispositions to feel pleasure and pain, are troublingly dynamic.

One early datapoint for preferential dynamics comes from Festinger & Carlsmith (1959). These psychologists divided participants into two groups, a high-payment group and a low-payment group, then asked them to complete a boring experimental task. After the task, participants were asked to pretend to other potential participants that the task was enjoyable. Months later, participants anonymously reported on how enjoyable

this boring task had been. Those in the high-payment group reported as expected that the task had been quite boring; surprisingly, however, those in the low-payment group reported that they had in fact enjoyed the task. Festinger & Carlsmith propose that this asymmetry can be explained in terms of cognitive dissonance. Participants found the task boring, but then they had to try to convince someone else that it was interesting. Deceiving in this way makes people uncomfortable, so they look for a way to reconcile their actual and stated opinions. Those in the high-payment group were able to rationalize that they had received sufficient compensation for the boringness of the task, but those in the low-payment group were not. Instead, they convinced themselves that they had in fact enjoyed the task, a process similar to the *sour grapes* phenomenon. In sour grapes, people who do not have something convince themselves that they do not want it. In this experiment, conversely, people who had something convinced themselves that they wanted it.

Though fascinating, the Festinger & Carlsmith (1959) study suffers from some serious methodological drawbacks. Preferences were elicited long after the event, allowing for forgetting, screen memories, and wishful thinking to do their dirty work. In addition, preferences were elicited through self-report rather than through action, leaving open the question whether participants in the low-payment condition were lying or deceiving themselves.

Later work in experimental economics, however, overcomes these defects. Hoeffler & Ariely (1999), for instance, demonstrated that when people are given a chance to explore a range of potential objects of preference, their preferences change

over the course of the exploration, though the changes get smaller as time goes by. This led to the development of the theory of arbitrary coherence, according to which indeterminate preferences become determinate over time through the making of choices. When we choose something, we more often than not end up preferring it, even though we would not have preferred it had we initially chosen something else. Thus, preferences start out indeterminate, go through a dynamic stage during which early choices largely determine later preferences and later choices, then finally ossify into stable preferences. If this is right, it clearly violates the principle of evaluative stasis, thereby spelling trouble for preference utilitarianism

In yet another experiment, Hoeffler, Ariely, & West (2006) found further evidence for both hedonic indeterminacy and hedonic dynamics. They gave participants lemonade of varying strengths – some clearly too weak, some clearly to strong, and several in between. Participants were then asked to find a recipe for what they considered the Goldilocks of lemonade. Those who had an initially unpleasant experience (too weak or too strong) searched more widely and systematically for the perfect recipe than did those who had an initially pleasant experience. The upshot is that what people find pleasurable depends in a troubling way on their early experiences. A given person's hedonic values start out indeterminate; then experiences cause them to stabilize around the cluster of their early experiences. But, of course, it seems obvious that early experiences need not be ideal or near-ideal.

In any event, even if the ossification of hedonic value around an arbitrary anchor is ultimately untroubling, it entails that hedonic utilitarianism violates the evaluative stasis principle.

10.1.4. Where do we go from here?

One obvious objection to the evidence for indeterminacy is to say that in fact the subjects *did* assign some value (preferential or hedonic) to the proposition in question, but that the valuation changed over the course of the experiment. Though I find this rebuttal dubious, I shall respond with a dilemma. Either the rebuttal is incorrect, in which case the indeterminacy argument stands unchallenged, or the rebuttal is correct, in which case its proponents have supplied further evidence for the argument from dynamic preferences, pleasures, and pains. The point remains, then, that both preference and pleasure- and pain-causation are subject to normatively irrelevant situational influences. Thus, any consequentialist theory that includes preference-satisfaction or pleasure- or pain-causation in its tables of value violates the evaluative stasis principle. The rebuttal is impaled on the twin horns of indeterminacy and dynamics.

If the experiences of defenders of virtue ethics are any guide, consequentialists should forgo futile attempts to refute the empirical challenge and instead focus on reshaping their theories in light of that challenge. I have argued elsewhere that the situationist challenge to virtue ethics should not be resisted but embraced by finding ways to leverage the power of situations to encourage behavior in accordance with virtue. Perhaps an embracement strategy would work for consequentialism as well.

Consider the following future-perfect accounts of goodness in terms of preference, pleasure and pain:

(10.9) An action is good if its consequences would be preferred after the action is performed.

(10.10) An action is good if its consequences would be found pleasurable after the action is performed.

While these suggestions obviously stand in need of further development (are consequences assessed immediately after, not long after, at the end of one's life, at the end of time?), I believe they offer a promising new way for consequentialists to define goodness, betterness, and rightness. It may entail a form of relativism, but a very interesting one according to which things get less relativistic the more experiences one has.

10.2. Future directions

10.2.1. Towards a new iconophilia

In the sixth century, Pope Gregory I wrote a letter (11:10) to an iconophobic English priest, reprimanding him for destroying images of saints. He admitted that icons could be used as idols, but attempted to rehabilitate them when used properly. In defense of iconophilia, Gregory argued that pictures of virtue teach their viewers to be virtuous. He contrasted the adoration of an image with learning from the image what one should adore, and went on to say that, in a picture, "even the ignorant see what they ought to follow; in it the illiterate read." Such a view stands in direct opposition to the view, articulated by Kupfer (2006, p. 341) but I suspect widely held, that "[e]xposure

to images of virtue alone is not likely to alter people much." Kupfer seems to think that people act virtuously when they have the right motivations and reasons. An image – even an image of virtue – does not provide either, so it is at best tangentially related to morality. Moral and social control is best exercised by giving people the right reason to act.

In 2009, the New York MTA began broadcasting the following message in subway cars: "If you see an elderly, pregnant, or handicap person near you offer your seat. You'll be standing up for what's right. Courtesy is contagious and it starts with you." This announcement epitomizes the reason-giving approach to situation-producerism: if people are given the right reason, they will act in accordance with it. Perhaps, though, something more subtle and seemingly irrelevant would work better. In section 9 I showed that people can be induced to behave more pro-socially by the tactical manipulation of social distance heuristics. Putting a dominant or trustworthy face in their field of vision leads them to underestimate social distance, and thence to behave more pro-socially. What if the MTA, instead of inflicting bad puns on straphangers, were to display images of faces with the optimal features for inducing pro-social behavior? Perhaps Gregory was right after all. Perhaps an aesthetically inclined philanthropist should donate the funds to find out.

10.2.2. Situationsim and the self

In closing, I would like to offer one further suggestion, this one on the nature of the self. The bounds of the situation are the bounds of the self. If the characterization of situations developed in sections 5 through 9 is correct, the self is more like an inflated

balloon than a solid ball. In those sections, I argued that the situation includes not only purely external features like ambient smells, temperature, and presence of bystanders, but also seemingly internal features like moods, emotions, and the like. Thus, the situation has both an external component and an internal component. The self is the boundary between these two: a membrane stretched between opposing pressures that together conspire to give it its shape.

Bibliography

Adams, R. (2006). *A Theory of Virtue*. Oxford: Oxford University Press.

Albarracín, D. & McNatt, P. (2005). Maintenance and decay of past behavior influences: Anchoring attitudes on beliefs following inconsistent actions. *Personality and Social Psychology Bulletin*, 31:6, 719-733.

Alfano, M. (forthcoming). Nietzsche, naturalism, and the tenacity of the intentional. *International Studies in Philosophy.*

Alfano, M. (2010). The tenacity of the intentional prior to the *Genealogy*. *The Journal of Nietzsche Studies*, 40, 123-140 (2010).

Allen, C. (1982). Self-perception based strategies for stimulated energy conservation. *The Journal of Consumer Research*, 8:4, 381-390.

Allport, G. (1966). Traits revisited. *American Psychologist*, 21, 1-10.

Andrade, E. B., & Ariely, D. (2009). The enduring impact of transient emotions on decision making. *Organizational Behavior and Human Decision Processes*, 109, 1-8.

Andreoni, J. & Petrie, R. (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics*, 88:7-8, 1605-1623.

Annas, J. (1993). *The Morality of Happiness*. New York: Oxford University Press.

Annas, J. (2003). Virtue ethics and social psychology. *A Priori*, 2, 20-59.

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy*, 33, 1-19.

Apsler, R. (1975). Effects of embarrassment on behavior toward others. *Journal of Personality and Social Psychology*, 32, 145-153.

Ariely, D. (2008). *Predictably Irrational*. New York: Harper Collins.

Ariely, D. & Norton, M. (2008). How action create – not just reveal – preferences. *Trends in Cognitive Science*, 12:1, 13-16.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118:1, 73-105.

Ariely, D., Loewenstein, G., & Prelec, D. (2006). Tom Sawyer and the construction of value. *Journal of Economic Behavior & Organization*, 60, 1-10.

Aristotle (2000). *Nicomachean Ethics*. Crisp (trans.) Cambridge: Cambridge University Press.

Athanassoulis, N. (2000). A response to Harman: Virtue ethics and character traits. *Proceedings of the Aristotelian Society*, 100, 215-222.

Audi, R. (2001). Epistemic virtue and justified belief. In *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*. Oxford: Oxford University Press.

Aumann, R. & Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica*, 63:5, 1161-1180.

Bacon, F. (1620). *Novum Organum*. Reprinted in E. A. Burtt, ed. (1939). *The English Philosophers from Bacon to Mill*. New York: Random House.

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons." *Evolution and Human Behavior*, 25, 209-220.

Baron, R. (1997). The sweet smell of … helping: Effects of pleasant ambient fragrance on prosocial behavior in shopping malls. *Personality and Social Psychology Bulletin*, 23, 498-503.

Baron, R. A., & Thomley, J. (1994). A whiff of reality: Positive affect as a potential mediator of the effects of pleasant fragrances on task performance and helping. *Environment and Behavior*, 26, 766-784.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, 12, 412-414.

Batson, C. (1991). *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Batson, C. (2002). Addressing the altruism question experimentally. In Post, Underwood, Schloss, & Hurlbut (eds.) *Altruism and Altruistic Love: Science, Philosophy, and Religion in Dialogue*, 89-105. Oxford: Oxford University Press.

Batson, C., Coke, J., Chard, F., Smith, D., & Taliaferro, A. (1979). Generality of the 'glow of goodwill': Effects of mood on helping and information acquisition. *Social Psychology Quarterly*, 42, 176-179.

Batson, C., van Lange, P., Ahmad, N., & Lishner, D. (2003). Altruism and helping behavior. In Hogg & Cooper (eds.) *The Sage Handbook of Social Psychology*, 279-295. London: Sage Publications.

Bentham, J. (1995). *The Panopticon Writings*. London: Verso.

Bentham, J. (1789/1961). *An Introduction to the Principles of Morals and Legislation*. Garden City: Doubleday. Originally published in 1789.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Bicchieri, C. & Chavez, A. (2009). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, forthcoming.

Bicchieri, C. & Xiao, E. (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, 22, 191-208.

Blass, T. (1999). The Milgram paradigm after 35 years: Some things we now know about obedience to authority. *Journal of Applied Social Psychology*, 29:5, 955-978.

Blum, L. (1994). *Moral Perception and Particularity*. Cambridge: Cambridge University Press.

Bohnet, I., & Frey, B., (1999a). Social distance and other-regarding behavior in dictator games. *The American Economic Review*, 89.1, 335-339.

Bohnet, I., & Frey, B., (1999b). The sound of silence in prisoner's dilemma and dictator games. *Journal of Economic Behavior and Organization*, 38, 43-57.

Boles, W. & Haywood, S. (1978). The effects of urban noise and sidewalk density upon pedestrian cooperation and tempo. *Journal of Social Psychology*, 104, 29-35.

Bowers, K. (1973). Situationism in psychology: An analysis and a critique. *Psychological Review*, 80, 307-336.

Brañas-Garza, P. (2007). Promoting helping behavior with framing in dictator games. *Journal of Economic Psychology*, 28, 477-486.

Brandt, R. (1970). Traits of character: A conceptual analysis. *American Philosophical Quarterly*, 7, 23-37.

Brandt, R. (1979). *A Theory of the Good and the Right*. New York: Oxford University Press.

Brandt, R. (1992). *Morality, Utilitarianism, and Rights*. Cambridge: Cambridge University Press.

Brennan, G., & Pettit, P. (1993). Hands invisible and intangible. *Synthese*, 94, 191-225.

Broome, J. (1991). *Weighing Goods*. Oxford: Basil Blackwell.

Buchan, N., Johnson, E., & Croson, R. (2006). Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preferences. *Journal of Economic Behavior and Organization*, 60:3, 373-398.

Burger, J. M., & Caldwell, D. C. (2003). The effects of monetary incentives and labeling on the foot-in-the-door effect: Evidence for a self-perception process. *Basic and Applied Social Psychology*, 25, 235-241.

Burnham, T. (2003). Engineering altruism: A theoretical and experimental investigation of anonymity and gift giving. *Journal of Economic Behavior and Organization*, 50, 133-144.

Burnham, T. & Hare, B. (2007). Engineering human cooperation. *Human Nature*, 18.2, 88-108.

Burnyeat, M. (1980). Aristotle on learning to be good. In A. Rorty (ed.) *Essays on Aristotle's Ethics*, 69-92. Berkeley: University of California Press.

Camerer, C. & Thaler, R. (1995). Anomalies: Ultimatums, dictators and manners. *Journal of Economic Perspectives*, 9, 209-219.

Campbell, J. (1999). Can philosophical accounts of altruism accommodate experimental data on helping behavior? *Australasian Journal of Philosophy*, 77, 26-45.

Carlsmith, J. & Gross, A. (1968). Some effects of guilt on compliance. *Journal of Personality and Social Psychology*, 53, 1178-1191.

Charness, G. & Gneezy, U. (2007). What's in a name? Anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior and Orgaization*, 63:1, 88-103.

Charness, G., Haruvy, E., & Sonsino, D. (2007). Social distance and reciprocity: The internet vs. the laboratory. *Journal of Economic Behavior and Organization*, 63:1, 88-103.

Chwe, M. (2001). *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton: Princeton University Press.

Cohen, S. (1978). Environmental load and the allocation of attention. In Baum, Singer, & Valins (eds.) *Advances in Environmental Psychology*, volume 1. Hillsdale, NJ: Erlbaum.

Cohen, S. & Lezak, A. (1977). Noise and inattentiveness to social cues. *Environment and Behavior*, 9, 559-572.

Cornelissen, G., Dewitte, S. & Warlop, L. (2007). Whatever people say I am that's what I am: Social labeling as a social marketing tool. *International Journal of Research in Marketing*, 24:4, 278-288.

Cornelissen, G., Dewitte, S., Warlop, L., Liegeois, A., Yzerbyt, V., Corneille, O. (2006). Free bumper stickers for a better future: The long term effect of the labeling technique. *Advances in Consumer Research*, 33, 284-285.

Crisp, R. (1996). Modern moral philosophy and the virtues, in Crisp, ed., *How Should One Live? Essays on the Virtues,* 1-18. Oxford: Oxford University Press.

Darley, J. & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, 27, 100-108.

Darley, J. & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8, 377-383.

Darwin, C. (1887/1958). *The Autobiography of Charles Darwin: 1809-1882*. Barlow (ed.). New York: Norton.

Dent, N. (1975). Virtues and actions. *The Philosophical Quarterly*, 25.

DesAutels, P. (2004). Moral mindfulness. In DesAutels & Urban Walker (eds), *Moral Psychology: Feminist Ethics and Social Theory*, Rowman and Littlefield, 69-81.

Diogenes Laertius. (1965). *Lives of Eminent Philosophers*. Volume 2. Translated by R D Hicks. Cambridge, Massachussetts: Harvard University Press.

Ditmarsch, H., Eijck, J., & Verbrugge, R. (2009). Common knowledge and common belief. In Eijck & Verbrugge (eds.) *Discourses on Social Software*, pp. 107-132. Amsterdam: Amsterdam University Press.

Donnerstein, E. & Wilson, D. (1976). Effects of noise and perceived control on ongoing and subsequent aggressive behavior. *Journal of Personality and Social Psychology*, 34, 774-781.

Doris, J. M. (1998). Persons, situations, and virtue ethics. *Nous*, 32:4, 504-540.

Doris, J. M. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Doris, J. M., & Stich, S. P. (2005). As a matter of fact: Empirical perspectives on ethics. In Jackson & Smith (eds.) *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Dorsey-Palmateer, R. & Smith, G. Regression to the mean in flight tests. Working Paper. Pomona College.

Driver, J. (2001). *Uneasy Virtue*. Cambridge: Cambridge University Press.

Eckel, C. & Grossman, P. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, 16, 181-191.

Epicurus. (1940). Fragments. In *The Stoic and Epicurean Philosophers*. Edited and translated by Whitney Oates. New York: Random House.

Epstein, S. (1983). Aggregation and beyond: Some basic issues in the prediction of behavior. *Journal of Personality*, 51, 360-391.

Feldman, F. (1997). *Utilitarianism, Hedonism, and Desert*. New York: Cambridge University Press.

Feldman, F. (2004). *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism*. New York: Oxford University Press.

Festinger, L., & Carlsmith, J. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal Social Psychology*, 58, 203-211.

Fields, J., & Schuman, H. (1976). Public beliefs about the beliefs of the public. *Public Opinion Quarterly*, 40, 427-448.

Fischbacher, U., Gächter, S., & Fehr, E. (forthcoming). Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters*.

Fischer, G. W., & Hawkins, S. A. (1993). Strategy compatibility, scale compatibility, and the prominence effect. *Journal of Experimental Psychology: Human Perception and Performance*, 19, 580-597.

Flanagan, O. (1991). *Varieties of Moral Personality*. Cambridge: Harvard University Press.

Fleming, D. (2009). The character of virtue: Answering the situationist challenge to virtue ethics. *Ratio*, 19:1, 24-42.

Foot, P. (1997). Virtues and vices, in Crisp & Slote (eds.) *Virtue Ethics*, 163-177. Oxford: Oxford University Press.

Foot, P. (2001). *Natural Goodness*. Oxford: Clarendon Press.

Foucault, M. (1997). *Discipline and Punish*. Sheridan (trans.). New York: Pantheon Books.

Fox, J. & Guyer, M. (1978). Public choice and cooperation in *n*-person prisoner's dilemma. *Journal of Conflict Resolution*, 22:3, 469-481.

Freedman, J. & Fraser (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, 4, 195-202.

Frey, B. S. & Bohnet, I. (1995). Institutions affect fairness: Experimental investigations. *Journal of Institutional and Theoretical Economics*, 151:2, 286-303.

Frey, B. S. & Bohnet, I. (1999a). Social distance and other-regarding behavior in dictator games. *American Economic Review*, 89:1, 335-339.

Frey, B. S. & Bohnet, I. (1999b). The sound of silence in prisoner's dilemma and dictator games. *Journal of Economic Behavior & Organization*, 38:1, 43-57.

Frischer, B. (1982). *The Sculpted Word*. Berkeley, CA: University of California Press.

Funder, D. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, 40, 21-34.

Funder, D. & Ozer, D. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology*, 44, 107-112.

Gächter, S. & Fehr, E. (1999). Collective action as a social exchange. *Journal of Economic Behavior and Organization*, 39:4, 341-369.

Geach, P. (1977). *The Virtues*. Cambridge: Cambridge University Press.

Geen, R. & O'Neal, E. (1969). Activation of cue-elicited aggression by general arousal. *Journal of Personality and Social Psychology*, 11, 289-292.

Gibbard, A. (1992). *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press.

Gigerenzer, G. (2007). *Gut Feelings*. New York: Viking.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (2000). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Goldman, A. (1993). Ethics and cognitive science. *Ethics*, 103, 337-360.

Gooodman, N. (1965). *Fact, Fiction, and Forecast*. New York: Bobbs-Merrill.

Green, D., Jacowitz, K., Kahneman, D., & McFadden, D. (1998). Referendum contingent valuation, anchoring, and willingness to pay for public goods. *Resources and Energy Economics*, 20, 85-116.

Gregory I. Letters. Translated by Adam McKeon.

Grimes, M.B. (1999). Helping behavior commitments in the presence of odors: Vanilla, lavender, and no odor. National Undergraduate Research Clearinghouse, 2.

Grusec, J. E., Kuczynski, L., Rushton, J. P., & Simutis, Z. M. (1979). Learning resistance to temptation through observation. *Developmental Psychology*, 15, 233-240.

Grusec, J. & Redler, E. (1980). Attribution, reinforcement, and altruism: A developmental analysis. *Developmental Psychology*, 16:5, 525-534.

Haley, K. & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26, 245-256.

Haney, C., Banks, W., & Zimbardo, P. (1973). Interpersonal dynamics of a simulated prison. *International Journal of Criminology and Penology*, 1, 69-97.

Harman, G. (1965). The inference to the best explanation. *The Philosophical Review*, 74:1, 88-95.

Harman, G. (1973). *Thought*. Princeton: Princeton University Press.

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, New Series 119, 316-331.

Harman, G. (2000). The nonexistence of character traits. *Proceedings of the Aristotelian Society*, 100, 223-226.

Harman, G. (2001). Virtue ethics without character traits, in Byrne, Stalnaker, & Wedgwood, eds., *Fact and Value*, 117-127. Cambridge: MIT Press.

Harman, G. (2003). No character or personality. *Business Ethics Quarterly*, 13:1, 87-94.

Harman, G. (2006). Three trends in moral and political philosophy. *The Journal of Value Inquiry*, 37.

Hartshorne, H., & May, M. (1928). *Studies in the Nature of Character*, volume 1. New York: Macmillan.

Hempel, C. (1966). Deductive-nomological vs. statistical explanation. In H. Feigl and G. Maxwell (eds.) *Minnesota Studies in Philosophy of Science*, vol. 3, 98-169. Minneapolis: University of Minnesota Press.

Henderlong, J. & Lepper, M. (2002). The effects of praise on children's intrinsic motivation: A review and synthesis. *Psychological Bulletin*, 128:5, 774-795.

Hodgman, J. (2001). Invisible man vs. hawkman. *This American Life*, 178.

Hoeffler, S. & Ariely, D. (1999). Constructing stable preferences: A look into dimensions of experience and their impact on preference stability. *Journal of Consumer Psychology*, 8:2, 113-139.

Hoeffler, S., Ariely, D., & West, P. (2006). Path dependent preferences: The role of early experience and biased search in preference development. *Organizational Behavior and Human Decision Processes*, 101, 215-229.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7, 346-380.

Hoffman, E., McCabe, K., Smith, V. (1996). Social distance and other-regarding behavior in dictator games. In Plott & Smith (eds.) *Handbook of Experimental Economics*, volume 1, 429-435. Amsterdam: Elsevier.

Hudson, S. (1980). Character traits and desires. *Ethics*, 90, 539-542.

Hurka, T. (2001). *Vice, Virtue, and Value*. Oxford: Oxford University Press.

Hurka, T. (2006). Virtuous act, virtuous dispositions. *Analysis*, 66, no. 1, 75.

Hursthouse, R. (1999). *On Virtue Ethics*. Oxford: Oxford University Press.

Isen, A. (1987). Positive affect, cognitive processes, and social behavior. In L. Berkowitz (ed.) *Advances in Experimental Social Psychology*, volume 20, 203-254. San Diego: Academic Press.

Isen, A., Clark, M., & Schwartz, M. (1976). Duration of the effect of good mood on helping: Footprints on the sands of time. *Journal of Personality and Social Psychology*, 34, 385-393.

Isen, A. M., & Levin, P. F. (1972). The effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology*, 21, 384-388.

Isen, A., Shalker, T., Clark, M., & Karp, L. (1978). Affect, accessibility of material in memory, and behavior: A cognitive loop. *Journal of Personality and Social Psychology*, 36, 1-12.

Jennings, D., Amabile, T., & Ross, L. (1982). Informal covariation assessment: Data-based vs. theory-based judgments. In Tversky, Kahneman, & Slovic (eds.) *Judgment Under Uncertainty: Heuristics and Biases*, 211-230. New York: Cambridge.

Jensen, A. & Moore, S. (1977). The effect of attribute statements on cooperativeness and competitiveness in school-age boys. *Child Development*, 48, 305-307.

Johnson, E., & Schkade, A. (1989). Bias in utility assessments: Further evidence and explanations. *Management Science*, 35, 406-424.

Jones, E. & Harris, V. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, 3, 1-24.

Jones, E. & Nisbett, R. E. (1971). *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*. New York: General Learning Press.

Jost, J. & Jost, L. (2009). Virtue ethics and the social psychology of character: Philosophical lessons from the person-situation debate. *Journal of Research in Personality*, 43:2, 253-254.

Kahneman, D., & Knetsch, J. (1993). Anchoring or shallow inferences: The effect of format. Unpublished manuscript, University of California, Berkeley.

Kamtekar, R. (2004). Situationism and virtue ethics on the content of our character. *Ethics*, 114, 458-491.

Keynes, J. M. (2009). *The General Theory of Employment, Interest, and Money*. New York: Classic Books America.

Khan, U. & Dhar, R. (2006). Licensing effect in consumer choice. *Journal of Marketing Research*, 43, 259-266.

Kilham, W. & Mann, L. (1974). Level of destructive obedience as a junction of transmitter and executant roles in the Milgram obedience paradigm. *Journal of Personality and Social Psychology*, 29, 696-702.

Konecni, V. (1975). The mediation of aggressive behavior: Arousal level versus anger and cognitive labeling. *Journal of Personality and Social Psychology*, 32, 706-716.

Korte, C. & Grant, R. (1980). Traffic noise, environmental awareness, and pedestrian behavior. *Environment and Behavior*, 12, 408-420.

Korte, C., Ypma, A., & Toppen, C. (1975). Helpfulness in Dutch society as a function of urbanization and environmental input level. *Journal of Personality and Social Psychology*, 32, 996-1003.

Kraut, R. (1973). Effects of social labeling on giving to charity. *Journal of Experimental Social Psychology*, 9:6, 551-562.

Kripke, S. (1972). *Naming and Necessity*. Cambridge: Harvard University Press.

Kristjansson, K. (2008). An Aristotelian critique of situationism. *Philosophy*, 83, 55-76.

Krupka, E., & Weber, R. (2006). The focusing and informational effects of norms on pro-social behavior. Working Paper, Carnegie Mellon University.

Kupfer, J. (2006). Film criticism and virtue theory. In Carroll & Choi (eds.) *Philosophy of Film and Motion Pictures: An Anthology*, 335-347. Oxford: Blackwell.

Kupperman, J. (1995). *Character*. Oxford: Oxford University Press.

Kupperman, J. (2001). The indispensability of character. *Philosophy*, 76, 239-250.

Kuran, T. & Sunstein, C. (1999). Availability cascades and risk regulation. *Stanford Law Review*, 51:4, 683-768.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, 28, 75-84.

Lakatos, I. (1995). *The Methodology of Scientific Research Programmes*. Worrall & Currie (eds.). Cambridge: Cambridge University Press.

Latané, B., & Darley, J. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology*, 10, 215-221.

Latané, B., & Darley, J. (1970). *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century-Crofts.

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, 89, 308-324.

Latané, B., & Rodin, J. (1969). A lady in distress: inhibiting effects of friends and strangers on bystander intervention. *Journal of Experimental Psychology, 5, 189-202.*

Levine, R., Norenzayan, & Philbrick, K. (2001). Cross-cultural differences in helping strangers. *Journal of Cross-Cultural Psychology*, 32, 543-560.

Lewis, D. (1986). Causal explanation. In *Philosophical Papers*, vol. 2, 214-240. Oxford: Oxford University Press.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89, 46-55.

Lichtenstein, S., & Slovic, P. (1973). Response-induced reversals of preference in gambling: An extension replication in Las Vegas. *Journal of Experimental Psychology*, 101, 16-20.

Lindbeck, A. (1997). Incentives and social norms in household behavior. *American Economic Review*, 87:2, 370-377.

Lipton, P. (2004). *Inference to the Best Explanation*, 2nd edition. New York: Routledge.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American Economic Review*, 90:2, 426-432.

MacIntyre, A. (1984). *After Virtue: A Study in Moral Theory*. Notre Dame: University of Notre Dame Press.

MacIntyre, A. (1997). The nature of the virtues. In Crisp & Slote (eds.) *Virtue Ethics*, 118-140. Oxford: Oxford University Press.

Mack, A. & Rock, I. (1998). *Inattentional Blindness.* Cambridge: MIT Press.

Malle, B. (2006). The actor-observer 194symmetry in causal attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132, 895-919.

Malle, B., Knobe, J., & Nelson, S. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology*, 93, 491-514.

Mantell, D. (1971). The potential for violence in Germany. *Journal of Social Issues*, 27, 101-112.

Marx, K. (1845/1998). *The German Ideology, including Theses on Feuerbach*. New York: Prometheus Books.

Marx, K. (1859/1904). *A Contribution to the Critique of Political Economy*. Kerr (trans.). New York: Macmillan.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2001). Monetary and non-monetary punishment in the voluntary contributions mechanism. Purdue University, Krannert Graduate School of Management, USA. Working Paper.

Matthews, K. E., & Cannon, L. K. (1975). Environmental noise level as a determinant of helping behavior. *Journal of Personality and Social Psychology*, 32, 571-577.

McDowell, J. (1979). Virtue and reason. *Monist*, 62, 331-350.

Merritt, M. (2000). Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice*, 3, 365-383.

Milgram, S. (1974). *Obedience to Authority*. New York: Harper and Row.

Mill, J. S. (1861/1998). *Utilitarianism*, edited with an introduction by Roger Crisp. New York: Oxford University Press.

Miller, C. (2003). Social psychology and virtue ethics. *The Journal of Ethics*, 7:4, 365-392.

Miller, C. (2009). Empathy, social psychology, and global helping traits. *Philosophical Studies*, 142:2, 247-275.

Miller, R., Brickman, P., & Bolen, D. (1975). Attribution versus persuasion as a means for modifying behavior. *Journal of Personality and Social Psychology*, 31:3, 430-441.

Mills, R. S., & Grusec, J. E. (1989). Cognitive, affective, and behavioral consequences of praising altruism. *Merrill-Palmer Quarterly, 35,* 299–326.

Mischel, W. (1968). *Personality and Assessment*. New York: Wiley.

Mischel, W. & Peake, P. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review*, 89, 730-755.

Mischel, W. & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing the invariances in personality and the role of situations. *Psychological Review*, 102:2, 246-268.

Monin, B. & Miller, D. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, 81:1, 33-43.

Moore, G. E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.

Moore, G. E. (1912). *Ethics*. Oxford: Oxford University Press.

Nussbaum, M. (1995). *Poetic Justice: The Literary Imagination and Public Life*. Boston: Beacon Press.

Nussbaum, M. (2000). *Women and Human Development: The Capabilities Approach*. New York: Cambridge University Press.

Oakley, J. (1996). Varieties of virtue ethics. *Ratio*, 9, 129.

Oosterhof, N. & Todorov, A. (2008). The functional basis of face evaluations. *Proceedings of the National Academy of Sciences of the USA*, 105, 11087-11092.

Ouellette, J., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin*, 124, 54-74.

Pacuit, E., Parikh, R., & Cogan, E. (2006). The logic of knowledge based obligation. *Synthese*, 149, 311-341.

Page, R. (1974). Noise and helping behavior. *Environment and Behavior*, 9, 311-334.

Parikh, R. (2002). Social software. *Synthese*, 132:3, 187-211.

Plato (1968). *The Republic of Plato*. Bloom (trans.). New York: Basic Books.

Plaza, J. (2007). Logics of public communications. *Synthese*, 158:2, 165-179.

Popper, K. (2002). *Conjectures and Refutations: The Growth of Scientific Knowledge*, 2nd edition. London: Routledge.

Prinz, J. (2009). The normativity challenge: Cultural psychology provides the real threat to virtue ethics. *The Journal of Philosophy*, 13:2-3, 117-144.

Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

Regan, J. (1971). Guilt, perceived injustice, and altruistic behavior. *Journal of Personality and Social Psychology*, 18, 124-132.

Rege, M. & Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics*, 88:7-8, 1625-1644.

Richter, G. (1965). *Portraits of the Greeks*. London: Phaidon Press.

Rigdon, M., Ishii, K., Watabe, M., & Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology*, 30.3, 358-367.

Ring, K., Wallston, K., & Corey, M. (1970). Mode of debriefing as a factor affecting subjective reactions to a Milgram-type obedience experiment: An ethical inquiry. *Representative Research in Social Psychology*, 1, 67-85.

Rosati, C. S. (1995). Persons, perspectives, and full information accounts of the good. *Ethics*, 105, 296-325.

Ross, L. (1977). The intuitive psychologist and his shortcomings. In Berkowitz (ed.) *Advances in Experimental Psychology*, vol. 10, 174-214. New York: Academic Press.

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

Ross, L. & Nisbett, R. E. (1991). *The Person and the Situation*. Philadelphia: Cambridge University Press.

Roth, A., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review*, 81, 1068-1095.

Russell, D. (2009). *Practical Intelligence and the Virtues*. Oxford: Oxford University Press.

Sabini, J. & Silver, M. (2005). Lack of character? Situationism critiqued. *Ethics*, 115, 535-562.

Sachdeva, S., Iliev, R., & Medin, D. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science*, 20:4, 523-528.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Sarkissian, H. (forthcoming). Minor tweaks, major payoffs: The problems and promise of situationism in moral philosophy. *Philosopher's Imprint*.

Schaller, M., & Cialdini, R. (1990). Happiness, sadness, and helping: A motivational integration. In Higgins & Sorrentino (eds.) *Handbook of Motivation and Cognition*, 265-296. New York: The Guilford Press.

Schelling, T. (1968). The life you save may be your own. In Chase (ed.) *Problems in Public Expenditure Analysis*, 127-162. Brooking Institute, Washington DC.

Schwartz, S. & Gottlieb, A. (1991). Bystander anonymity and reactions to emergencies. *Journal of Personality and Social Psychology*, 39, 418-430.

Scott, C. & Yalch, R. (1980). Consumer response to initial product trial: A Bayesian analysis. *The Journal of Consumer Research*, 7, 32-41.

Sen, A. (1979). Utilitarianism and welfarism. *Journal of Philosophy*, 76, 463-89.

Sen, A. (1985). Well-being, agency, and freedom. *Journal of Philosophy*, 82:4, 169-221.

Seneca. *Letters to Lucilius*. Translated by Richard Sorabji.

Shargel, D. Unpublished Ph.D. Dissertation. Emotions as bodily states.

Sheridan, C. & King, R. (1972). Obedience to authority with an authentic victim. *Proceedings of the eightieth annual convention of the American Psychological Association*, 165-166. Washington DC: American Psychological Association.

Sidgwick, H. (1907). *The Methods of Ethics*. London: Macmillan.

Simmel, G. (1964). *The Sociology of Georg Simmel*. Wolff (trans.). Glencoe: Free Press.

Singer, P. (1993). *Practical Ethics, Second Edition*. Cambridge: Cambridge University Press.

Sinnott-Armstrong, W. P., (2005). Moral intuitionism meets empirical psychology. In Horgan & Timmons (eds.) *Metaethics after Moore*. New York: Oxford University Press.

Sizer, L. (2000). Towards a computational theory of mood. *The British Journal for the Philosophy of Science*, 51:4, 743-770.

Slote, M. (1992). *From Morality to Virtue*. Oxford: Oxford University Press.

Slovic, P. (1995). The construction of preference. *American Psychology*, 50, 364-371.

Smith, M. (1995). Internal reasons. *Philosophy and Phenomenological Research*, 55.

Snow, N. (2009). *Virtue as Social Intelligence: An Empirically Grounded Theory*. New York: Routledge.

Solomon, R. (2003). Victims of circumstances? A defense of virtue ethics in business. *Business Ethics Quarterly*, 13, 43-62.

Solomon, R. (2005). What's character got to do with it? *Philosophy and Phenomenological Research*, 71, 648-655.

Solomon, R. (forthcoming). Comments on John Doris' *Lack of Character*. *Philosophy and Phenomenological Research*.

Spinoza, B. (1677/1992). *Ethics, Treatise on the Emendation of the Intellect, and Selected Letters*. Shirley (trans.) and Feldman (ed.). Indianapolis: Hackett.

Sreenivasan, G. (2002). Errors about errors: Virtue theory and trait attribution. *Mind*, 111, 47-68.

Sreenivasan, G. (2008). Character and consistency: Still more errors. *Mind*, 117, 603-612.

Stocker, M. (1976). The schizophrenia of modern moral theories. *Journal of Philosophy*, 73.

Strawson, P. F. (1960). Freedom and resentment. *Proceedings of the British Academy*, 48, 1-25.

Strenta, A. & DeJong, W. (1981). The effect of a prosocial label on helping behavior. *Social Psychology Quarterly*, 44:2, 142-147.

Sunstein, C. (2005). Moral heuristics. *Behavioral and Brain Sciences*, 28, 531-542.

Swanton, C. (2003). *Virtue Ethics: A Pluralist View*. Oxford: Oxford University Press.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, 5:223, 79-97.

Tajfel, H. (1973). The roots of prejudice: Cognitive aspects. In Watson (ed.) *Psychology and Race*. Penguin.

Tajfel, H. (1981). *Human Groups and Social Categories*. Cambridge: Cambridge University Press.

Tajfel, H. (1982). Social psychology and intergroup relations. *Annual Review of Psychology*, 33, 1-30.

Tannsjo, T. (1998). *Hedonistic Utilitarianism*. Edinburgh: Edinburgh University Press.

Taylor, R. (1991). *Virtue Ethics*. Interlaken, New York: Linden Books.

Thomson, J. J. (1996). Evaluatives and directives, in Harman & Thomson, eds., *Moral Relativism and Moral Objectivity.* Oxford: Blackwell.

Thomson, J. J. (1997). The right and the good. *Journal of Philosophy*, 94, no. 6, 280-283.

Tronto, J. (1993). *Moral Boundaries: A Political Argument for an Ethic of Care*. New York: Routledge.

Tversky, A., & Kahnemann, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5, 207-232

Tybout, A. & Yalch, R. (1980). The effect of experience: A matter of salience? *Journal of Consumer Research*, 6, 406-413.

Uleman, J. et al. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inference. In Zanna (ed.) *Advances in Experimental Social Psychology, volume 28*, 211-280. San Diego: Academic Press.

Upton, C. (2005). A contextual account of character traits. *Philosophical Studies*, 122, 133-151.

Upton, C. (2009a). The structure of character. *The Journal of Ethics*, 13, 175-193.

Upton, C. (2009b). Virtue ethics and moral psychology: The situationism debate. *The Journal of Ethics*, 13, 103-115.

Vaidyanathan, R. & Praveen, A. (2005). Using commitments to drive consistency: Enhancing the effectiveness of cause-related marketing communications. *Journal of Marketing Communications*, 11:4, 231-246.

Van Rooy, R. (2003). Quality and quantity of information exchange. *Journal of Logic, Language, and Information*, 12, 423-451.

Vranas, P. (2005). The indeterminacy paradox: character evalutions and human psychology. *Nous*, 39, 1-42.

Wallace, J. (1974). Excellences and merit. *The Philosophical Review*, 83:2, 182-199.

Wallace, J. (1978). *Virtues and Vices*. Ithaca: Cornell University Press.

Walton, G., & Spencer, S. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of stereotyped students. *Psychological Science*, 20:9, 1132-1139.

Watson, G. (1990). On the primacy of character. In Flanagan & Rorty (eds.) *Identity, Character, and Morality: Essays in Moral Psychology*, 449-483. Cambridge: MIT Press.

Webber, J. (2006a). Character, consistency, and classification. *Mind*, 115.

Webber, J. (2006b). Virtue, character and situation. *Journal of Moral Philosophy*, 3:2, 193-213.

Webber, J. (2007a). Character, common-sense, and expertise. *Ethical Theory and Moral Practice*.

Webber, J. (2007b). Character, global and local. *Utilitas*.

Wedekind, C. & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current Biology*, 12, 1012-1015.

Weyant, J. (1978). Effects of mood states, costs, and benefits on helping. *Journal of Personality and Social Psychology*, 36, 1169-1176.

Wielenberg, E. (2006). Saving character. *Ethical Theory and Moral Practice*, 9, 461-491.

Williams, B. (1981). *Moral Luck*. Cambridge: Cambridge University Press.

Williams, B. (1985). *Ethics and the Limits of Philosophy*. Cambridge: Harvard University Press.

Williams, B. (1998). Virtues and vices. In Craig (ed.) *Routledge Encyclopedia of Philosophy*. London: Routledge.

Winter, M. & Tauer, J. (2006). Virtue theory and social psychology. *Journal of Value Inquiry*, 40, 73-82.

Yzerbyt, V. et al. (2001). The dispositional inference strikes back: Situational focus and dispositional suppression in causal attribution. *Journal of Personality and Social Psychology*, 81, 365-376.