THE CHARACTER OF CHARACTER: NEW DIRECTIONS FOR A DISPOSITIONAL THEORY

Nicole Smith

Submitted to the Graduate College of Bowling Green State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

August 2013

Committee:
Daniel Jacobson, Advisor
Judith A. Zimmerman, Graduate Faculty Representative
Christian Coons
David Shoemaker
ABSTRACT

Daniel Jacobson, Advisor

My dissertation aims to solve a puzzle, a paradox, and a problem. The puzzle is to explain why people act in uncharacteristic (i.e., seemingly cruel) ways in a number of social psychological experiments, such as Stanley Milgram’s obedience experiment, in which 65% of the participants complied with the experimenter’s demands to issue a series of increasingly powerful “shocks” to an unwilling recipient. I argue that owing to features of the experimental design participants were made to feel: out of their element, confused, disoriented, pressured, intimidated, and acutely distressed, and that the “experimenter” (actually a confederate) exploited these factors, which is the central reason why the majority of participants complied with his demands despite being reluctant to do so.

The paradox is that, although ordinary people seem to be good, bad, or somewhere in between, evidence (again from social psychology) seems to suggest that most people would behave deplorably on many occasions and heroically on many others. This, in turn, suggests the paradoxical conclusion that most people are indeterminate—i.e., no particular character evaluation appears to apply to them. I argue to the contrary that the social psychological evidence fails to support the claim that people would behave deplorably on many occasions. Milgram’s participants, for example, faced extenuating circumstances that should mitigate the degree to which they were blameworthy for their actions, and this, in turn, challenges the claim that they behaved deplorably.
The problem is that no existing theory has been able to adequately account for the connection between possessing certain character traits and performing certain actions. Commonsense suggests that there is a connection between, for instance, being a truthful person and telling the truth, but it has been challenging for philosophers to capture precisely what the connection is in an empirically defensible way. I argue that there is a strong empirical connection between character traits and action, but that to understand this connection, it must be admitted at the outset that interfering factors may come between traits and their manifestations in action. The key is to develop a theory that can give an account of which sources of interference are character trait undermining and which are not. I go some way toward developing such a theory.
To Ian: For encouraging me to go through, 
    not around the hard problems. 
And I’m so pleased Ms. Bronte got it right!

To my mom, one of the best moral psychologists I know: 
    For helping me to appreciate 
    the value of perseverance.

To Dan, Dave, and Christian: 
    For helping me over the rainbow and back.

In loving memory of my dad (Here’s to hoisting up the John B. sails!) 
    and 
Shiloh (“The coolest beagle in Ann Arbor”).
ACKNOWLEDGMENTS

I am deeply grateful to Bill Simpson for his support and guidance during my years as an undergraduate—without it, I’m fairly certain I would have lost my way. A special thanks is also owed to Dan Jacobson, David Shoemaker, Christian Coons, and Sarah Buss for their support over the last several years. I would also like to thank Mark Alfano, Nomy Arpaly, Emily Austin, Neera Bahdwar, Eric Brown, John Brunero, Ray Buchanan, Randy Clarke, Sinan Dogramaci, John Doris, Julia Driver, Billy Dunaway, Jerry Green, Alex Grzankowski, Chris Heathwood, Joshua Hollowell, Charlie Kurth, Steve Finlay, Heidi Maibom, Hugh Marlowe, Michael Mckenna, Christian Miller, Howard Nye, Jason Raibley, Alex Plakias, Adam Pautz, Ian Proops, Rachelle Robles, Mark Sainsbury, Mark Schroeder, Tim Schroeder, Michael Smith, Daniel Star, Trisha Trigilio, Eric Vogelstein, Arthur Ward, Eric Wiland, Susan Wolf, and Judith Zimmerman. In addition, I want to acknowledge the members of the audiences at the North Carolina Philosophical Society Meeting and the Rocky Mountain Ethics Congress in 2011. I am grateful to the philosophy department at Bowling Green State University for awarding me a much needed dissertation fellowship from 2009–2010, in which the seeds of this project were sown.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER I. BULLIES IN LAB COATS</td>
<td>10</td>
</tr>
<tr>
<td>1 Background</td>
<td>11</td>
</tr>
<tr>
<td>2 Obedience</td>
<td>13</td>
</tr>
<tr>
<td>3 A Bully in Experimenter’s Clothing</td>
<td>19</td>
</tr>
<tr>
<td>4 Implications for Moral Psychology</td>
<td>37</td>
</tr>
<tr>
<td>5 Conclusion</td>
<td>39</td>
</tr>
<tr>
<td>CHAPTER II. GOOD PEOPLE, BAD CIRCUMSTANCES</td>
<td>41</td>
</tr>
<tr>
<td>1 The Indeterminacy Paradox</td>
<td>42</td>
</tr>
<tr>
<td>2 The Fragmentation Thesis</td>
<td>45</td>
</tr>
<tr>
<td>3 The Stanford Prison Experiment</td>
<td>45</td>
</tr>
<tr>
<td>4 The Milgram Experiment</td>
<td>48</td>
</tr>
<tr>
<td>5 Against the Fragmentation Thesis</td>
<td>51</td>
</tr>
<tr>
<td>6 Objections and Replies</td>
<td>71</td>
</tr>
<tr>
<td>7 The Goodness of Ordinary People</td>
<td>78</td>
</tr>
<tr>
<td>CHAPTER III. FAILURES TO MANIFEST OR MANIFEST FAILURES?</td>
<td>86</td>
</tr>
<tr>
<td>1 The Trait-Action Problem</td>
<td>89</td>
</tr>
<tr>
<td>2 The Dispositional Theory of Character Traits</td>
<td>90</td>
</tr>
<tr>
<td>3 The Masking Thesis</td>
<td>97</td>
</tr>
<tr>
<td>4 The Empirical Thesis</td>
<td>101</td>
</tr>
</tbody>
</table>
INTRODUCTION

Concepts related to a person’s character feature prominently in our ordinary discourse and thought: some people strike us as sweet and kind, while others come across as complete jerks. History tells us about some downright evil people we wish had never been born, and, yet, we can take some comfort in contemplating the lives of the occasional moral saint. We often evaluate people in terms of their character (e.g., we say they are good or bad) and we also attribute character traits to them (e.g., we say they are conscientious, loyal, or generous). We are inclined, moreover, to think of many of those we encounter in our daily lives as generally decent people. In general, concepts of character, broadly construed, serve as a compass by which we navigate a substantial portion of the normative terrain. Indeed, we often reference these features of persons when we hold them responsible or when we deem them praiseworthy or blameworthy for their actions, or even when we excuse or forgive them for their mistakes.

The trouble is that the application of character concepts is strained in contexts where people behave in seemingly uncharacteristic ways. How, for instance, can we make sense of someone who normally would not hurt a fly, but who nonetheless finds himself in a circumstance where he issues what he takes to be increasingly painful shocks to an unwilling recipient? The aim of my dissertation is to solve this and other puzzles, and to propose new directions for theorizing about character traits.

BEYOND THE STANDARD INTERPRETATIONS

Moral psychology has taken on new life in recent years owing to the rapid expansion of interdisciplinary research in philosophy, psychology, neuroscience, evolutionary biology, among
others. In particular, psychology, of both the social and cognitive varieties, is quickly becoming deeply integrated into normative inquiry, and nowhere is this more apparent than in work on character. I welcome this development, for there is no better place to start in the explanation of human action and behavior than with detailed descriptions of the actual actions and behaviors of ordinary people in real (or controlled) situations. It leaves less to the philosopher’s imagination —avoiding what Nomy Arpaly has referred to as “an unbalanced diet of examples” (2003: 20)— and relies more heavily on the very subject matter at issue; namely, moral agents in action. As with any new development, however, caution is recommended, especially when drawing substantive philosophical conclusions from abstract data or (sometimes) oversimplifications of the causal influences underlying the behavior of experimental subjects.

One area in need of improvement is how psychologists portray their experiments to a popular audience. Social psychology, in particular, has managed to produce no shortage of headlines over the last half-century. The growing acknowledgment stems not just from their “shocking” results—an adjective frequently used in connection with Milgram’s early work—but also from grandiose analogies. For instance, Milgram (1963; 1974) compared his subjects with perpetrators of the Holocaust. Latané and Darley (1970) drew comparisons between their work on passive bystanders and those who witnessed (and were reported to) do nothing about the murder of Kitty Genovese—a claim made in a widely publicized news story that has since been shown to be unsubstantiated (Manning et al. 2007). Philip Zimbardo claims that his 1971 Stanford Prison Experiment helps us to understand how good people turn evil (2007: 5–6). He thinks his experimental paradigm can help us understand the actions of those involved in the the Rwanda genocide. I think these analogies are not only overblown, but dangerous and
intellectually irresponsible. Furthermore, they encourage distorted caricatures of the experiments, obscuring the underlying factors necessary for providing an adequate explanation of the participants’ actions.

Part of the task at hand is to go beyond the metaphors, caricatures, and standard interpretations, more generally. Consider what is sometimes simply referred to as “The Milgram Experiment.” Milgram conducted not one, but over nineteen variations of the experiment. The famous “baseline” condition, in which 65% of subjects issued a series of “shocks,” up to and including 450 volts, to an unwilling recipient in response to the experimenter’s demands, is one among roughly half of the experiments where the majority of participants “obeyed.” In the others, the majority of participants resisted. This is relevant for several reasons. It is a reminder, first of all, that the circumstances must be fairly unique in order for most people to behave in the kinds of problematic ways observed in the baseline condition. Also of interest here is the recent archival work describing Milgram’s exploratory methods (Russell 2011). It turns out that Milgram had altered the experimental design in various ways plausibly because he thought it would help secure the surprising result we have all come to know so well.

Milgram frames his studies as revelatory of obedience to authority, but what is almost never mentioned is that the experimenter abused his authority by pressuring participants to issue shocks beyond the point where they were no longer willing to issue them. I argue at length in the first chapter that the experimenter’s tactics constituted a kind of bullying, and that, once we understand how this affected participants’, their actions are more readily explained.

One defect of standard interpretations is a failure to engage in projective empathy with the participants, by which I mean, to try to get a sense of what it is was like from their
perspective. It is an easy temptation to succumb to, for few of us would want to be in their shoes.
Nevertheless, I argue that once we try to entertain what the participants in the experiment could have been thinking and feeling, we will find that the degree to which the participants are blameworthy for their actions should be mitigated somewhat because of the difficult circumstances they faced. As for Zimbardo’s Stanford Prison experiment, I think a similar lesson should be drawn once we learn that much of the humiliating and degrading treatment of the “prisoners” (in fact college students) was built into the design of the experiment, not always perpetrated by the “prison guards” (also college students).

In general, social psychologists would do better to maintain higher standards of rigor in the process of designing and conducting their experiments as well as in the reporting of the results. For instance, the heavy emphasis placed on publishing only positive or “exciting” results has come under fire recently as has the failure to replicate results.

Much of the negative attention has focused on various “priming” studies conducted by cognitive psychologists (Bartlett 2013), but social psychology has its own share of problems. For instance, Isen and Levin’s (1972) frequently cited experiment on the purported effects of mood on helping behavior (mood elevation is said to be positively correlated with helping behavior, while the absence of mood elevation is not) has had an unsuccessful record of replication, as other authors have noted (Miller 2009). I speculate that this is because Isen and Levin failed to control for factors, such as the “body language” of the person in need of help. Whether it is appropriate to help in non-emergency situations in part depends on whether the person seems to want help, and, yet they do not report having controlled for this factor. I discuss this experiment, and a related one by Matthews and Canon (1975), in more detail in chapter three. In any case,
one of the new directions I take in the following chapters is to move beyond standard interpretations of the relevant social psychological experiments.

BEYOND CHARACTER-BASED ETHICAL THEORY

Traditionally, the concept of character has been intimately bound up with the concept of human excellence and with the question of “How to live?” I am sympathetic to this general approach as long as it is recognized that the path to excellence may very well be paved with stones that are cracked, missing, or displaced in ways that are liable to trip one up along the way. As Robert Adams says, “it is important to find moral excellence in imperfect human lives” (2006: 119). In my view, philosophers doing moral psychology must face up to this simple fact before theorizing about any normative issue. And, yet, it is occasionally suggested that the concept of character does not even apply to those who have not yet reached the upper bound of human excellence. The tendency to draw this conclusion is most prominently found in the work of philosophers concerned with character-based ethical theories, especially those in which Aristotle’s phronimos—the practically wise moral expert—features prominently. While the emphasis on such an idealized figure with perfect character may aid in proposing an account of right action, such a theory is not well-placed to adequately account for the kind of character traits that apply to ordinary people.

The problem, as I see it, is that on the most plausible theory available, a person possesses a character trait partly in virtue of possessing dispositions to act in certain ways in certain circumstances, but these dispositions need not manifest to be possessed, so it is unclear from the start what the connection is supposed to be with respect to possessing character traits (including
virtues) and performing right actions. This problem has pushed virtue theorists in two opposing— and I contend equally problematic— directions. The first is to add a rider to the account of character traits that specifies a degree of frequency for which the trait-relevant dispositions must manifest. Adams (2006) defends a version of this view for some traits, and I address its problems in the third chapter.

The second direction virtue theorists can go is to focus exclusively on “inner states,” such as the person’s motives or, alternatively, the agent’s faculty of practical reason, each of which may or may not translate into overt behavior. The central problem with the “inner state” view is that it simply gives up on the goal of specifying what the connection between traits and actions is, even though commonsense would suggest that there is such a connection. The trouble with emphasizing the role of practical reason is that it blocks a central theoretical advantage of thinking about character traits in dispositional terms; namely, dispositions can explain how someone might be morally meritorious for performing an action without first deliberating or reflecting via one’s faculty of practical reason. This is because dispositions appear to be reason-responsive in that they are activated by the person’s awareness of a certain state of affairs (e.g., that someone needs help), and if that disposition leads all the way to action (she helps), then, all else being equal, she is praiseworthy for performing that action (helping) regardless of her failure to deliberate. Virtue ethicists who place heavy emphasis on practical reason must, it seems, forfeit this advantage of the dispositional account.

The last point I want to make in this connection is that character-based ethical theorists seem to err in their theoretical priorities. A theory of character traits should be articulated and defended prior to, and independent of, questions about the role these traits may or may not play
in ethical theory. Only then can we ensure a non-\textit{ad hoc} theory of character traits that is plausible on general theoretical and conceptual grounds, as well as satisfying certain explanatory desiderata. In the chapters that follow, I aim to make some progress toward this end.

\section*{BEYOND SITUATIONISM}

One recent author, Mark Alfano, describes the situationist challenge to virtue ethics in the following way: “if people are unconsciously susceptible to such seemingly trivial and normatively irrelevant influences as their degree of hurry, receiving cookies, and finding dimes, one can only infer that they would also be swayed by major temptations” (Alfano 2013: 37). Perhaps in the end, Alfano’s remarks do, in fact, suggest a legitimate critique of virtue ethics. I will not hazard a guess here. But, situationists, such as John Doris and Gilbert Harman, often claim that their arguments work equally well against \textit{anyone} who posits the existence of character traits, and it is this claim I take issue with in the chapters that follow.

Are people unconsciously susceptible to trivial and normatively irrelevant influences? I argue throughout that, although initially puzzling, the actions (or omissions) of the participants in many of the experiments situationists reference can be explained, and what appear to be trivial and normatively irrelevant influences are, in fact, normatively significant. For instance, situationists tend to follow the psychologists, Matthews and Canon (1975), in thinking that it is the aversiveness of the loud noise of a power mower that correlates with a failure to help someone pick up their items. But this would not explain why most people fail to help even the absence of the loud noise. In the third chapter, I argue that whether someone helps in circumstances such as this has more to do with the presence or absence of certain forms of
communication, such as whether the person in need of help conveys welcoming “body language.” These, sometimes tacit, forms of communication are hardly normatively irrelevant. Furthermore, explicit forms of communication are obstructed by the loud noise, so it is not clear this factor is doing the primary explanatory work.

To their credit, situationists correctly put their finger on one problem with current theories of character traits, and this is the failure of those theories to accommodate the complexity of human moral life. I depart from situationists, however, in thinking that in order to accommodate this complexity we need to substantially revise our understanding of character traits. In the third chapter, I suggest that the concepts that will aid the proponent of a dispositional theory of character traits are already found in ordinary discourse. We are sometimes inclined to say that a certain person “acted out of character,” and we do not mean to imply that they now lack character. Similarly, a person might possess a character trait, while also having certain kinds of “character flaws” that interfere with it. Accordingly, I recommend a shift in emphasis from theorizing about character traits as “statistical facts,” as Brandt would say (1970: 34), where the focus is on the number of times such traits fail to manifest in action and behavior, and, instead, turn our attention to the nature of these failures, specifically whether they undermine or preserve our character trait attributions on a given a occasion.

One further reason I suggest we move in a different direction from the one recommended by situationists is twofold: first, their criticism of dispositional theories of character traits presuppose a certain version of that view which had already been thoroughly criticized and rejected on metaphysical and conceptual grounds by Ryle (1949) and Brandt (1970) so many
years earlier. Accordingly, neglecting these arguments has resulted in a serious obstacle to progress on the topic.

The second, and related problem is, in my view, the general resistance on the part of situationists to engage with the metaphysical literature. A concrete example of this is found in John Doris’s monograph (2002). Here, he begins by correctly noting that the analysis of dispositions in terms of a simple counterfactual conditional has been disputed. On the face of it, this is a problem for the proponent of a dispositional theory of character traits, but instead of acknowledging it as such, Doris has this to say: “Such examples seem to show that the conditional analysis fails. No bother: I’m doing moral psychology, not metaphysics; my interest is not in conceptual analysis but in the evidential standards governing trait attribution” (2002: 16). In this passage, Doris tries to change the subject from metaphysical and conceptual issues to epistemic ones, but this does not stop him from making metaphysical claims about the nature of character or about how many people possess the dispositional properties associated with character traits.

The lesson seems to be that there is no escaping these issues, there are only temporary methods for evading them. In the chapters that follow I attempt to confront these problems head on. In chapter three, in particular, I defend an empirically sensitive and metaphysically informed understanding of the dispositional theory of character traits.
CHAPTER 1: BULLIES IN LAB COATS: NEW IMPLICATIONS OF THE MILGRAM EXPERIMENT FOR MORAL PSYCHOLOGY

In a recent interview, Bob McDonough—whose father played the role of the “victim” in a number of Milgram’s obedience experiments—reported a remarkable coincidence. His neighbor, a New Haven firefighter, had been one of the participants who had issued “shocks” to his father during one of the experiments. And, yet, “in an odd twist of events,” McDonough added, “when my father collapsed in the kitchen the morning he died in January 1965, my sisters ran across the street and got this same guy to come over and administer mouth to mouth” (Davis and Perry 2008: 17). One question McDonough’s anecdote leaves us contemplating is how an ordinary guy—a firefighter who regularly puts his life on the line to help others—could comply with demands that would have him issue a series of increasingly painful “shocks” to an unwilling victim. The explanatory puzzle to which Milgram’s obedience experiments give rise is how to make sense of the mismatch between the participants’ actions during the experiment and their actions in everyday life. In what follows, I endeavor to unravel this puzzle.

Despite sustained effort on the part of philosophers and psychologists to understand the so-called “obedience” experiments, I argue that standard proposals remain unsatisfactory. In my view, this is because a crucial piece of the puzzle is missing: namely, the human face of the experiments. If we scrutinize the details, attending to the reactions of the participants both during and after the experiment, it becomes clear that, owing to certain peculiarities of the circumstance, participants were made to feel: out of their element, confused, disoriented, pressured, intimidated, and acutely distressed. I argue, further, that the “experimenter” (actually a confederate) used tactics that exploited their confusion, disorientation, and distress, and that this
is the central reason why the majority of participants complied with his demands despite being unwilling to do so.

In the first section of the paper, I provide some context for the various permutations and results of the obedience experiments. In the second, I discuss the dangers of viewing these experiments through the distorting lens Milgram provides; namely, the Nazi Holocaust. I argue that obedience to authority is not fundamental to the explanation of the participants’ actions in the Milgram experiments. In the third section, I defend my account against the proposals of John Sabini and Maury Silver (2005), on the one hand, and Maria Merritt, John Doris, and Gilbert Harman (2010), on the other. In the fourth section, I draw out some implications of the Milgram experiment for moral psychology. In particular, I argue that, contrary to the claims of some philosophers, the participants’ actions in no way constitute evidence for the frailty of human moral agents.

I begin by offering a rather lengthy description of the experiment. Such detail may seem unnecessary, but, in the sections that follow, I aim to illustrate why it is necessary to go even further, beyond the superficial details of the experiment.

1. BACKGROUND

In the early 1960’s, Stanley Milgram conducted a series of experiments designed to illuminate the phenomenon of obedience. The basic structure of each experiment was the same.\(^1\) Participants responded to an advertisement requesting volunteers—factory workers, businessmen, construction workers, clerks, etc.—for “a study on memory and learning” (1963:

\(^1\) The description that follows relies heavily on the discussions in Milgram 1963: 371–376 and Milgram 1974: 13–43).
When they arrived at the designated location, however, the volunteers learned an additional fact: the goal of the study was specifically to test the effects of punishment on memory and learning. Subsequently, each participant was assigned a role—either “teacher” or “learner.” The process of selecting these roles appeared random to the participant, but was, in fact, rigged—the confederate always played the learner. The teacher and learner then took up their respective positions: First, the teacher watched as the learner was strapped into a chair and attached with electrodes, and then the experimenter brought the teacher to a different room and positioned him in front of the “shock generator.” Here the experimenter gave the learner a “sample shock” of 45-volts in order to fortify his belief in the authenticity of the experiment.

The teacher’s instructions were to issue a shock to the learner every time he failed to remember the correct word pair. The cleverly designed, although nonfunctioning, shock generator had small levers corresponding to thirty levels of “voltage” ranging from 15 to 450. Once the experiment commenced, the experimenter (also a confederate) issued the following sequence of scripted “prods” each time the teacher hesitated: Prod 1: “Please continue,” or alternatively, “Please go on.” Prod 2: “The experiment requires that you continue.” Prod 3: “It is absolutely essential that you continue.” Prod 4: “You have no other choice, you must go on” (Milgram 1974: 21). The result was that 65% of the forty subjects who took part in the original experiment completed it by administering every “shock” level up to and including 450-volts (Milgram 1974: 35). In total, Milgram conducted nineteen variations on the original experiment (Milgram 1974: 119; Davis and Perry 2008: 14). He varied factors such as: the distance between the experimenter and the teacher, the distance between the teacher and the learner, the location, the gender of the participants, and the actors who played the confederates.
The focus of this paper will primarily be on the fifth variant of the experiment, what Milgram called the “new baseline condition,” as well as its relatives, the first and second experiments.\footnote{The difference between them is minimal. In the first experiment, the learner was non-verbal, but the teacher could hear him pounding through the wall. In the second, the teacher could hear the learner’s protests. The fifth experiment adds additional dialogue to the learner’s script. Among his protests, is the complaint that he is experiencing heart trouble (Milgram 1974: 56).}

The participants in these experiments were all adult males (not students from the university). Levels of compliance were roughly the same in all three experiments. Where relevant, I discuss some of the other experimental conditions as well (I indicate explicitly to this effect on such occasions).

\section*{2. OBEDIENCE}

Milgram’s aim in designing the experiment was to better understand the phenomenon of obedience. He saw this issue as particularly important given the actions committed by the military and citizenry of Nazi Germany. In this connection, he says:

It has been reliably established that from 1933 to 1945 millions of innocent people were systematically slaughtered on command. Gas chambers were built, death camps were guarded, daily quotas of corpses were produced with the same efficiency as the manufacture of appliances. These inhumane policies may have originated in the mind of a single person, but they could only have been carried out on a massive scale if a very large number of people obeyed orders (1974: 1).

Many have followed Milgram in thinking that his experiments are important to understanding obedience and the atrocities committed in its name.\footnote{More recently, Peter Vranas (2005) and Jesse Prinz (2009) draw explicit comparisons between the Milgram experiments and the Holocaust. One obvious dissimilarity is that, unlike the Nazi’s attitudes toward non-Arians, Milgram’s participants did not possess malicious, contemptuous, or other hateful attitudes toward the learner.} A reviewer from the \textit{Washington Post}
describes Milgram’s book, *Obedience to Authority*, as “the classic account of the tendency to follow orders, no matter who they hurt or what their consequences” (Milgram 1974 [2009]: front cover). In a review of a different book, the philosopher, Peter Singer, says, “Milgram’s experiments on obedience have made us more aware of the dangers of uncritically accepting authority” (Milgram 1974 [2009]: back cover; Singer 2004). This raises the question, to what extent were Milgram’s participants simply taking orders. In what follows, I argue that obedience to authority is not primarily what is at issue in the explanation of the actions and behaviors of the participants in these experiments.

2.1 WERE THEY “JUST FOLLOWING ORDERS”?

We know the experimenter issued instructions and demanded participants comply with them. Is this why they complied? Were they “just following orders”? If such an explanation were correct, it would be fairly surprising given that participants had a compelling reason not to comply with the experimenter’s demands; namely, complying would require them to do something morally impermissible. Accordingly, advocates of this way of understanding the Milgram experiment owe us an explanation of why participants would comply with the experimenter’s unjustified demands.⁴ There are several possibilities here.

First, a subject might comply with an unjustified demand if he fails to realize that it is unjustified. Sabini and Silver’s explanation lends itself to this interpretation in certain places. They claim that the experimenter played the role of “institutional expert” insofar as he

---

⁴ A demand can be unjustified in two ways: first, obeying the demand might require one to do something morally impermissible. Second, the act of issuing the demand might be morally unjustified. In the case in question, however, the act of issuing the demand is morally unjustified because the demand requires one to do something that is morally impermissible. Accordingly, there is no reason to disambiguate the two notions here.
“interpreted for the subject” which action was the most appropriate one in the circumstance 
(2005: 550; emphasis added). They elaborate further: “the experimenter is a guide to this little bit 
of the social/moral world. He is taken to be someone who sees clearly what is objectively 
needed” (2005: 551). Importantly, in their view, the experimenter suggests that issuing the 
shocks is the right thing to do and the subjects believe him. Were this interpretation correct, it 
would explain why the majority of participants complied with the experimenter’s morally 
unjustified demands: on this view, they simply did not think the demands were unjustified.

The difficulty for this interpretation, however, is that evidence from the experiment 
suggests that, once the shock levels reached a certain level, the majority of subjects no longer 
believed that issuing the shocks was morally permissible, and, so, did not think the demands 
justified. As Milgram says, “subjects often expressed deep disapproval of shocking a man in the 
face of his objections, and others denounced it as stupid and senseless” (1963: 376).
Furthermore, at various stages of the experiment, most participants hesitated, expressed their 
desire to end the experiment, and/or exhibited signs of severe stress and frustration (Milgram 
1965; 1963: 375). Were it the case that participants considered the experimenter’s demands to be 
justified, then, assuming they were cooperative subjects, we would expect uniform compliance, 
without hesitation, through to the end of the experiment. But this is not what was observed.

Another reason why someone might comply with an unjustified demand is if that person 
possed the disposition to unquestioningly obey authority. Such a person would have a 
tendency to comply with the authority’s demands even when they were unjustified (and regarded 
as such by the subject). Here I agree with Lee Ross and Richard Nisbett in thinking it
implausible that the majority of subjects in the Milgram experiment possessed this disposition. They say:

We do not find evidence that people are disposed to obey authority figures unquestioningly—even to the point of committing harmful and dangerous acts. (We’ve never found such slavish obedience forthcoming from the students or advisees whom we enjoin to keep up to date with their reading, take neat lecture notes, and to study for exams in an orderly and timely fashion—and it is not because they do not regard us as authority figures.) (1991: 59).

Furthermore, it would be antecedently unlikely that Milgram’s subjects—all of whom lived in America, an anti-authoritarian culture—would comply with the demands of the experimenter merely because they possessed a disposition to unquestioningly obey an authority.5

Third, and finally, someone might comply with an unjustified demand when the threat of sanction or punishment attends non-compliance. To illustrate, a soldier who disobeys orders might expect to be dishonorably discharged, imprisoned, or worse (if we are imagining a trial conducted within a morally unjust institution). Thus, we can understand why someone in this kind of circumstance might feel impelled to comply with an unjustified demand. In the Milgram experiments, however, there is no threat of sanction or punishment for non-compliance. The worst the experimenter can do is to dismiss the participant without paying them their four dollars—a compensation so negligible it can hardly count as a sanction or punishment. It is clear, moreover, that many participants were willing to forgo the monetary compensation so that they might be dismissed from the experiment (Milgram 1963: 376; Milgram 1965).

5 By contrast, it is not at all implausible to think that a young SS officer, indoctrinated from the age of thirteen as member of the Hitler Youth movement would cultivate dispositions of unquestioning obedience. This point further speaks to the danger of framing the Milgram experiment as shedding light on the kind of obedience prevalent in Nazi Germany.
We have yet to find a good reason for thinking that Milgram’s participants were simply following orders. For, if this proposal is to be at all plausible, it must explain why participants would comply with unjustified demands. But none of the explanations considered so far apply to Milgram’s participants. Of course, the experimenter is an authority and the majority participants did obey his demands. But I want to suggest that they did not comply for the sake of obedience simply because he was an authority. We have to look elsewhere for an explanation of why participants would comply with the experimenter’s unjustified demands.

2.2 AN ALTERNATIVE

Set aside the Milgram experiment for a moment. If one searches through possible explanations for why and in what contexts one is inclined to comply with someone’s demands—even though one takes those demands to be unreasonable or unjustified—the thought that comes immediately to my mind is that this kind of thing happens when someone has bullied, pressured, or intimidated me into complying. I will argue that, contrary to any prima facie hesitations one might have about applying this explanation to the Milgram experiment, once we look at the details—and in particular, the experimenter’s behavior as well as the reactions of the participants both during and after the experiment—the explanation begins to look quite plausible. Notice, moreover, that if we accept this explanation, then we should no longer see the Milgram experiment as merely an illustration of the power of authority to induce obedience. The reason is simple: obedience is an activity that involves either willingly or unwillingly complying with a demand, but successful bullying secures unwilling compliance with a demand. The Milgram
experiment, if I’m right, is an illustration of the power of bullying to bring about compliance with unjustified demands.

Before diving into the details of the experiment, I want to warn against a possible misinterpretation of my position. My appeal to the phenomenon of bullying may at first seem disconcertingly apologetic. After all, most people are appalled when they learn the results of the original Milgram experiment. A standard reaction is to think that the participants were seriously blameworthy for what they did (or would have done had they actually issued the shocks). Furthermore, most people think they would never have themselves done such a thing in like circumstances. For someone in this frame of mind, the claim that participants were bullied looks like an attempt to excuse the subjects’ actions. This is not, however, the use to which I intend to put the hypothesis. I am not arguing that Milgram’s subjects were justified in issuing the shocks. Nor am I addressing the issue of their moral responsibility. Rather, my aim here is purely explanatory.

A second concern must be allayed before we go on. The claim that Milgram’s subjects were bullied may appear to be at odds with certain facts about the case. After all, the experimenter’s manner as he demanded compliance could not have been more calm and cool. This aspect of the experimenter’s comportment is often emphasized in descriptions of the experiment. Merritt, Doris, and Harman, for example, suggest that the participants complied “at the polite request” of the experimenter (2010: 357). Or, consider Hagop Sarkissian’s description: “subjects were willing to administer shocks [in response to] the gentle prodding of the experimenter” (2010: 2; emphasis added). Finally, consider what Bill Menold—a participant in

---

6 Milgram conducted several surveys in support of this last point (1974: 27–31).
one of the initial experiments—said in an interview many years later: “the question was always geez, what can they make you do here, you know? Or what did you do? They didn’t make you. No one held a gun to my head” (Davis and Perry 2008: 13). Of course, a bully need not resort to threats to get what he wants; his tactics may be far more nuanced. My interpretation can therefore accommodate the fact that so few people found the experimenter’s behavior to be out of the ordinary. The simple fact is that the experimenter’s bullying tactics were rather subtle.7

3. A BULLY IN EXPERIMENTER’S CLOTHING

In this section, I argue that participants complied with the experimenter’s unjustified demands in part because they were made to feel:

(1) out their element;
(2) confused and disoriented;
(3) pressured;
(4) intimidated; and,
(5) acutely distressed.

I then defend what I take to be the best explanation for why participants felt these ways, comparing and contrasting it, where relevant, with the explanation offered by Sabini and Silver (2005), on the one hand, and Merritt, Doris, and Harman (2010), on the other. The most salient difference between the explanation defended here and those of the aforementioned authors is the role the experimenter plays in explaining the participants’ compliance. I argue that the experimenter used tactics that exploited their vulnerable state of mind, and that this was how he

---

7 One could argue that in order to be a genuine bully, the person in question must have certain malicious intentions. I disagree, but nothing hinges on the point here. My argument only requires that the experimenter was the functional equivalent of a bully.
managed to get the majority of them to comply with his demands against their will.

3.1 NOT IN KANSAS ANYMORE

In the original baseline condition of the experiment, unsuspecting participants arrived at “the elegant interaction laboratory” on the Yale campus (Milgram 1963: 372). They were ushered briefly through introductions, to find themselves confronted with the kind of thing one expects to see only in the movies. The learner was strapped into an “electric chair” and electrodes were then applied with electrode paste, which the experimenter announces, is to prevent blisters or burns (Davis and Perry 2008: 7). Bill Menold recollects how he watched as the learner loosened his tie and shirt, rolled up his sleeves, and was strapped into the chair, at which point he thought to himself “Holy mackerel, what is going on here?” (Davis and Perry 2008: 7). The participants were then seated face-to-face with the ominous “shock machine.” David Baker, the director of the archives of the history of American Psychology, describes his reaction to encountering it for the first time:

The thing that struck me and for anyone seeing it for the first time in person is really how large it is. It’s rather imposing with a black metal case, with a grey steel panel, with a series of switches and lights that go from 15 volts to 450 volts, and, then, underneath in black and red ink is the degree of shock that was supposedly being delivered [e.g., very strong shock] (Davis and Perry 2008: 5).

---

8 The location per se did not have pronounced effects, as the data from two other experimental conditions suggests. Milgram found that whether it was in the interaction laboratory, or a basement location on the Yale campus, or a “shady” office in Bridgeport, the average number of compliant subjects remained roughly the same (Milgram 1974: 66–70). We should not be too quick, however, to say that location was entirely irrelevant for all the subjects. In the Bridgeport condition, for instance, two people opted not to issue any shocks at all (Milgram 1974: 61). This did not occur in either of the other two conditions.
Menold—who was one of the fully compliant participants—said, “I was so disturbed by all this because [it] had gone out of my realm of reality and I was in a bizarre environment and I didn’t know what I was doing” (Davis and Perry 2008: 2). Owing to Milgram’s meticulous attention to detail in this elaborately staged experiment, participants found themselves completely out of their element and taken in by the illusion.

These features of the experimental design, I want to suggest, contribute to the explanation of the widespread compliance found among participants. When a person is in an unfamiliar place, confronting novel circumstances, he is less likely to act appropriately in the face of whatever surprising developments may occur in this unexplored territory. To some extent, this is because a person’s dispositions are finely tuned to the nuances of everyday life, not to the vicissitudes of unfathomable circumstances. Milgram’s participants were not even remotely prepared for what was about to happen, and the laboratory setting was a constant reminder of this fact.

3.2 DAZED AND CONFUSED

Two primary factors contributed to the confusion and disorientation participants felt in the course of the experiment. First, they were ignorant about several crucial facts with respect to their circumstances. For instance, unless the participant happened to have experience with wiring and electricity (and we can expect that few of them did), they did not know exactly how dangerous the shocks were; nor did they know the precise degree of pain the shocks would cause. Furthermore, once the experiment was in progress, and questions about the safety of the learner began to arise, participants were uncertain whether the shocks could aggravate the
learner’s (stated) heart condition; whether the experimenter who refused to check in on him could be trusted; and, whether the learner was unconscious or dead when he stopped responding. Finally, participants were almost certainly bewildered by conflicting pieces of information. For example, the experimenter reassured participants that the shocks “were painful but not dangerous,” yet, at around 375-volts, the shock generator reads “Danger: Severe Shock” (Milgram 1965; 1974: 28). Thus, participants had no shortage of misleading, dubious, and conflicting information to come to grips with in the heat of the moment. All of these factors, it is plausible to suppose, contributed to their state of confusion.

The second main factor contributing to subjects’ confusion and disorientation was the experimenter’s unexpected behavior. We enter any social situation expecting the people we are interacting with to comport themselves in a reasonable manner. But Milgram’s participants found themselves dealing with an experimenter who would not listen to reason; who would not take “no” for an answer; and, whose responses were almost entirely inappropriate to the concerns they had raised.

To illustrate, consider the following actual (and typical) exchange between the experimenter and one of the participants, as documented in Milgram’s film of the experiment, *Obedience* (1965) (“T” will stand for “teacher” and “E” will stand for “experimenter”):

[Several shocks have been issued. The learner has begun to yelp in pain and declares that he no longer wants to be a part of the experiment.]

T: [Turns to the experimenter] Do you want...? Do you want me to uh...?

E: Continue please, go right ahead.

T: But he refused to go on.
E: The experiment requires you to continue, teacher. Please continue. Next word is “sad.”

T: Want me to keep goin’?

E: Continue please.

[Next question. Wrong answer. Shock issued.]

T: That guy’s hollerin’ in there.

E: Continue please. Go on.

T: [Sighs deeply] [Mutters something] ...he might have a heart condition. Do you want me to go?

E: Just continue please.

[Next question. Wrong answer.]

T: [Addresses learner] Wrong! [Addresses experimenter] I’m up to 180 Volts. [Addresses learner] ...you’re going to get a shock of 180 volts. [Issues shock. Learner cries out in pain.] He can’t stand it. I’m not going to kill that man in there. You hear him hollerin’ in there?

E: As I said before, the shocks may be painful, but they’re not dangerous.

T: [Talks over the experimenter] Yeah, but he’s in there hollerin’. He can’t stand it. What happens if something happens to him?

E: The experiment requires that you continue, teacher.

T: Yeah, but...huh, I’m not going to get that man sick in there. Look...it..., he’s hollerin’ in there. You know what I mean, huh...[mumbles]. [Experimenter interrupts].
E: Whether the learner likes it or not, we must go on until he learns all the word pairs.

T: [Shakes his head] I refuse to take the responsibility of him gettin’ hurt in there. I mean I’m not goin’ to... [breaks off]. He’s in there hollerin’ in there.

E: It’s absolutely essential you continue teacher.

T: There’s too many left here. I mean, geez. He gets it wrong here. There’s too many of them left. I mean who’s going to take the responsibility if anything happens to that gentleman?

E: I’m responsible if anything happens here. Continue Please.

T: [Pauses] Alright. [sighs] Next one...

(It should be noted that the participant had only issued shocks up to 180-volts at this point, and he is one of the participants who eventually got to 450-volts. In fact, he went on to issue eighteen more shocks and had many more exchanges of this kind with the experimenter.) Now, a reasonable experimenter would have said, “Yes, you’re right, the learner appears to be in pain. We’d better check in on him.” Or, “It sounds like the learner is no longer willing to continue the experiment, so I’m afraid we’ll have to terminate it.” If the experimenter had said something to this effect, I imagine the participants would have thought him quite reasonable and been rather relieved. But, again, the experimenter did nothing of this sort; his behavior was utterly unreasonable.

This is partly why I have reservations about Sabini and Silver’s characterization of the experimenter’s role. They argue that the experimenter achieves compliance by functioning as the moral and institutional expert in the situation:
He “tells” them that the right thing to do here is to shock the victim. He does this not by arguing explicitly that it is morally correct to continue shocking but simply by ordering them in the most matter-of-fact way to continue. In so doing, he shows them that he (a seemingly reasonable, smart, competent fellow) takes it to be morally appropriate to do so (Sabini and Silver 2005: 551).

On their interpretation, participants complied because they trusted the experimenter—they believed (with him) that issuing the shocks was the right thing to do. Although, I am willing to agree that participants took the experimenter’s word for it on certain issues (I discuss examples shortly), this cannot be the whole story. After all, participants appear to be in two minds about whether to trust the experimenter. Evidence for this comes from exchanges such as the one above, and others like it. They reveal ambivalence and dubiousness, on the part of the subjects, about the appropriateness of the experimenter’s demands, especially once the learner began to protest. It is certainly true that the experimenter should be reasonable given his role as the institutional authority. It is also the case that participants expect him to be reasonable, but he defies their expectations, and in doing so creates a state of confusion and disorientation that escalates as the experiment progresses.9

There is an element of truth, however, to Sabini and Silver’s proposal. It seems that participants did believe the experimenter when he said, “I’m legally responsible for anything that happens” (Davis and Perry 2008: 11; cf. Milgram 1965). After all, participants seemed somewhat reassured by the experimenter’s remarks to this effect, and often they would press on in response (Milgram 1974: 160). The reason participants are moved by this consideration, in my view, is not

9 In this connection, it is worth noting that the experimenter abused his authority by demanding that participants act in morally impermissible ways. He went beyond the purview of what he could legitimately demand of them.
simply because they are selfishly concerned about the negative consequences for them if the learner is injured (or worse). Rather, a natural thought for them to have at this juncture is that the experimenter would not have them do anything that would legally jeopardize either the experimenter or the university, so the subjects continue issuing the shocks comforted by the thought that the learner has not been, and will not be, seriously injured as a result.

If this is right, then it appears that subjects were also compartmentalizing to a significant extent. For they seem to also think that the shocks may be causing the learner severe pain, and that this, too, is a reason to refrain from issuing more shocks, independent of whether they are injuring the learner. But somehow this thought is not salient to them at this particular moment of choice. The question then becomes, why are they compartmentalizing? A plausible answer, that ties in with the discussion above, is that participants were overwhelmed in their attempt to reconcile so much disparate and conflicting information. The situation seems amiss, but participants cannot quite put their finger on why. One of the participants, Bill Menold, said that, when the learner fell silent toward the end of the experiment, he had the following thought:

One of three things is happening: either this guy’s unconscious, he’s dead, or this thing is a complete sham. I mean I had thought, while my thought process was still working, [...] that I was being set up because you don’t do this—this isn’t the way the world operates. But the conflict within me to know which one of those was right was unbelievably stressful (Davis and Perry 2008: 11).

Accordingly, in the confusion and disorientation of the moment, participants lost not only their grip on reality, but their moral bearings as well.
The participants’ state of mind, in this regard, therefore contributes to the explanation of widespread compliance in two ways. First, when a person is caught off-guard by someone’s unexpected behavior, it much more difficult to think clearly about how best to confront him. Second, once a person is in a confused and disoriented state of mind, he is more susceptible to being persuaded to do things he is unwilling to do.

3.3 UNDER PRESSURE

It is rarely, if ever, noted in the philosophical literature that the experimenter pressured participants to comply. Nevertheless, when we bring the details of the experiment into sharper focus, it becomes clear that this was in fact the case. How did this come about? Well, one familiar and effective way we can be pressured by someone into acting against our will is by their refusing to take “no” for an answer. It is clear from the exchange above that this is precisely the situation participants faced. The experimenter inundated them with demands and he was unresponsive to their expressions of reluctance, as well as to the reasons they gave for not wanting to continue.

Participants also had every reason to think that the experimenter would not relent in his demands. In reality, of course, this was not the case. They were excused if they resisted after the experimenter issued four consecutive “prods” (1963: 374). We should be careful, however, not to be misled into thinking that it was easy for participants to meet the requirement for being excused. As observers, we are in the privileged position of knowing exactly what was required of them. By contrast, participants were utterly in the dark. The experimenter’s responses provided no indication that anything the participant said was going to persuade him to terminate the
experiment. In fact, in the post-experimental interview, Milgram asked the participant from the above exchange, “Why didn’t you just stop?” The participant replies anxiously, raising his voice: “He wouldn’t let me stop! I kept insisting to stop, but [the experimenter] says, “no”” (Milgram 1965). The many reactions of the participants in the experiment should leave little doubt that participants felt pressured to comply, and that the source of this pressure was owed, in part, to the content of the demands and the way in which the experimenter issued them.

Related to this point, the experimenter issued his demands rapidly, sometimes interrupting the participant, and never allowing him time to gather his thoughts. As Milgram says, “the experiment gives the subject little time for reflection” (1963: 378). This is relevant because participants were in no position to consider alternatives to compliance, and many of them failed to see what was, from an observer’s perspective, the obvious way out. Bill Menold says: “The thought of quitting.... It’s really strange. It never occurred to me to just say, “You know what, I’m walkin’ out of here,” which I could have done” (Davis and Perry 2008: 12). The experimenter dictates the pace, ensuring the momentum pushes the participant in the direction of compliance.

The pressure the experimenter exerted contributes to the explanation of the widespread compliance found among Milgram’s participants in a fairly straightforward way. We all possess limitations on our willpower, so when someone pressures or browbeats us into doing something we do not want to do, it is often because at some point we find ourselves unable to resist their demands.10 Milgram’s participants were just as susceptible to these tactics as any of us would be.

---

10 There are various explanations available for why we are susceptible to this phenomenon, including “ego depletion,” see Mark Muraven and Roy F. Baumeister 2000, and for a different take, see Veronica Job, et. al. 2010.
3.4 INTIMIDATION

Evidence suggests that participants also felt intimidated by the experimenter. Joseph Dimow—a participant in one of the initial experiments—recollects his experience:

I think at that time I’d had enough and I said that I wanted to stop...And [the experimenter] stood up and stood over me and started telling me how important the experiment was and that I was wasting their time and his time if I refused to continue with it (Davis and Perry 2008: 10).

It is possible, of course, that Dimow’s recollection of the exact words and events is not entirely reliable. Nevertheless, it would be implausible to suggest that he is mistaken about how the experimenter’s demeanor made him feel.\(^\text{11}\)

The question remains: why did participants feel intimidated? Aside from the natural connection between intimidation and pressure of the kind discussed above, Dimow’s recollection provides another clue. The physical proximity of someone can be a source of intimidation. We use phrases, such as “She got in my face” or “He was up in my grill,” to indicate that the person was being confrontational.

As it turns out, moreover, the data from other experimental conditions supports this proposal. Milgram discovered that the closer the experimenter was to the participant, the more likely it was that they would comply. He did this by comparing the results of the baseline condition, where the experimenter was in the room, to a condition where the experimenter issues instructions by telephone, to a further condition, in which the room was equipped with a tape recorder of the experimenter providing instructions. He reported:

\(^\text{11}\) One might be tempted into thinking that Dimow is confabulating this story in order to make himself look better, but this would implausible because he doesn’t need to make himself look better, he was not one of the fully compliant participants. He defied the experimenter (Davis and Perry 2008: 10).
Obedience dropped sharply as the experimenter was physically removed from the laboratory. The number of obedient subjects when the experimenter was present was almost three times as great as when the experimenter gave his instructions by telephone. It would appear that something akin to fields of force diminishing in effectiveness with increasing psychological distance from their source have a controlling effect on the subject’s performance (Milgram 1965; emphasis added).

Although intriguing, Milgram’s metaphor, “fields of force,” does not help to explain the correlation between, on the one hand, the proximity of the experimenter to the participant, and, on the other hand, the likelihood that the participant will comply. On the contrary, his appeal to a metaphor, in this context, illustrates a surprising obliviousness about typical human social interactions and the importance of tacit forms of communication, such as body language and physical proximity.

Interestingly, neither Sabini and Silver nor Merritt, Doris, and Harman do much better than Milgram in accounting for the explanatory significance of the experimenter’s proximity to the participants. Sabini and Silver are completely silent on the issue of physical proximity and compliance. Merritt, Doris and Harman go one step further. They claim that the physical proximity of the experimenter mattered: it caused participants to experience a certain kind of distress (a topic I discuss shortly), and this in turn, partially explains why the majority of participants complied. Yet, they never address the question why the physical proximity of the experimenter causes these participants to undergo distress. The explanatory question simply gets pushed back a step.
My account of these matters is one that should ring true with most people’s experiences. The physical proximity of a person can be intimidating. And when a person feels intimidated by someone, they are more susceptible to being persuaded to do things despite their reluctance to do them. Milgram’s participants, I have been arguing, are affected in this way by the experimenter, which is why intimidation via physical proximity contributes to the explanation of widespread compliance.

3.5 ESCALATING DISTRESS RESPONSES

The data suggest that most participants experienced distress during the course of the experiment (Milgram 1974: 42). In a telling description, someone who had witnessed the experiment reported the following:

I observed a mature and initially poised businessman enter the laboratory smiling and confident. Within 20 minutes he was reduced to a twitching, stuttering wreck, who was rapidly approaching a point of nervous collapse. He constantly pulled on his earlobe, and twisted his hands. At one point he pushed his fist into his forehead and muttered: “Oh God, let’s stop it” (Milgram 1963: 377).

Reactions of this sort were not uncommon among the participants (Milgram 1963: 375). Perhaps this is why recent discussions of the experiment have gone some way toward trying to account for this aspect of the participants’ behavior.

For their part, Merritt, Doris, and Harman claim that subjects were undergoing “personal distress,” which they stipulate is “a feeling of aversion to the target’s standpoint and affective state” (2010: 379). This is important because personal distress is incompatible with having
empathy for the learner, i.e., taking his perspective and sharing congruent emotions with him, and, so, ultimately incompatible with developing sympathy for him—“a benevolent sense of fellow-feeling” (2010: 379). Accordingly, on their view, if participants were undergoing personal distress, they would be less likely to come to the learner’s aid and more likely to comply with the experimenter’s demands.

The problem that arises for this interpretation of the experiment is that it is not clear whether Merritt, Doris, and Harman are correct in thinking that the participants were undergoing personal distress—in their stipulated sense of the term—for the duration of the experiment even though it is clear that subjects were undergoing distress of some kind. To illustrate the phenomenon of personal distress in more detail, consider someone who has just witnessed an extremely violent and horrific car crash. Instead of thinking about what sort of pain and trauma the victims are undergoing, the person might instead be occupied only by his own affective responses. Such a person may be awe-struck, terrified, and disturbed (quite understandably) in a way that prevents him from taking up the perspective of the victims. He may run away or freeze-up, but we can expect that until the personal distress passes, he is not likely to be rushing to anyone’s rescue. Is this what Milgram’s participants were experiencing?

As already indicated, if participants were undergoing personal distress, then they would also be unable to empathize with the learner, and, so, be incapable of feeling sympathy for him. Yet, typical exchanges between the participant and the experimenter tell a different story. Although participants never came to the rescue of the learner, they did, nonetheless, express concern for him at various stages of the experiment. They also worried about the level of pain he was experiencing, as well as the condition of his heart. In one exchange, not unlike the one
transcribed above, the participant says “I’m getting to the point where I can feel each [shock] with him” (Davis and Perry 2008: 9). Evidence therefore suggests, that we cannot be sure that the kind of distress they were undergoing at every stage of the experiment is fundamentally incompatible with their having empathy and, ultimately, sympathy for the learner.

If Merritt, Doris, and Harman have erred on the side of positing too extreme a form of distress to account for the behavior of the participants, then Sabini and Silver have erred in the opposing direction by focusing only on the milder forms. In fact, they do not address the topic of stress directly, but they do discuss motivational factors that could play an important role in explaining why participants experienced some low-level stress, as we will see.

Let’s assume that most participants felt uncomfortable issuing even the lowest level shocks. If this is right, and I think it is, then complying with the experimenter’s instructions even at the early stages of the experiment might be expected to have resulted in some distress. Part of what makes this situation stressful is that participants would have been naturally hesitant to decline cooperation at a point when, by their own lights, no moral transgressions had occurred. Had they opted out of issuing the shocks at this stage, they would have satisfied one of their preferences, but this would have been counterbalanced by the fact that doing so—and now I am drawing on the remarks of Sabini and Silver—would have required a rather awkward and embarrassing confrontation with the experimenter—not to mention having them violate several norms of politeness. Participants, therefore, may have experienced some distress in the early
stages of the experiment because they could not satisfy their preference to issue no shocks, while
at the same time avoiding an awkward confrontation with the experimenter.\footnote{I grant, along with Sabini and Silver, that the anticipation of embarrassment is a powerful motive. A startling example of this comes from a recent interview on This American Life. A woman explains how, as a 13-year-old, she had been attacked by a shark and was lying on the beach bleeding profusely when she turned down assistance from some fisherman passing by because her swimming suit had been mangled during the attack leaving her partially exposed. (They, of course, helped her anyway). This example reveals that we may be willing to undergo extreme costs to our own wellbeing to avoid embarrassment, but the question remains whether we can be brought to threaten other people’s wellbeing for the sake of avoiding embarrassment. I see no reason to think we can. For another example of the power of embarrassment, see Milgram’s discussion of his subway experiments in an interview with Carol Tavris (1974: 72).}

But this cannot be the whole story. For, what about the distress participants felt at other
stages of the experiment? Milgram says that participants “were observed to sweat, tremble,
stutter, bite their lips, groan, and dig their fingernails into their flesh,” and, much more
alarmingly, in the original experiment: “full-blown, uncontrollable seizures were observed for 3
subjects” (1963: 375). A suggestion that flows naturally from what both sets of authors have
suggested is that participants experienced different levels of distress at various stages in the
experiment. It seems natural to think that fully compliant subjects would have run the gamut
from mild to moderate to extreme distress. If this is right, then a plausible interpretation of the
experiment must account not only for the various levels of distress participants experienced, but
it must also be able to explain the source of the distress, and also why it became more
pronounced as the experiment progressed. In what follows, I intend to fill these explanatory
gaps.

In addition to the mild and extreme forms of distress described above, participants also
experienced moderate, but no less significant distress. The crucial point at which their level of
distress became elevated is surprisingly precise—or so I want to claim: it occurred just after they
issued the 150-volt shock. Two important developments happened at this point. First, the learner
began to protest, declaring that he no longer wanted participate in the experiment. At lower
levels he had let out an audible yelp, but at the 150-volt mark the participant could hear him cry and, apparently in pain, say: “Experimenter! That’s all, Get me out of here. I told you I had heart trouble. My heart’s starting to bother me now. Get me out of here, please. My heart’s starting to bother me. I refuse to go on. Let me out” (Milgram 1974: 56). It is at this point, I want to argue, that participants transitioned from being uncomfortable in issuing the shocks to being unwilling to do so. This is indicated by their verbal behavior (as illustrated in the extended transcript above) and also by the fact that compliance fell off sharply at the 150-volt mark (Milgram 1974: 56). It is reasonable to speculate, moreover, that this was the point at which participants judged that issuing further shocks would be morally problematic.

The second relevant development that occurred at this stage is that participants were no longer deterred by the prospect of an awkward confrontation with the experimenter. We know this for the simple reason that this is when they began to question the experimenter’s demands. Importantly, moreover, it is their resistance that provoked the onslaught of bullying tactics detailed above. Thus, at the 150-volt mark, the situation is fundamentally different: participants now want no part of the experiment (almost certainly because they think that continuing the experiment would be morally wrong), and, yet, at the same time the experimenter is insisting that they continue. This predicament, one can imagine, would indeed be stressful.

It is worth noting that the notion of distress appealed to in the above explanation is compatible with one of the prevailing theories of stress in the psychological literature. On Richard Lazarus and Susan Folkman’s (1984) theory, when a person cannot cope with the demands of a situation, i.e., when he judges that he lacks the relevant resources to meet the perceived demands of the situation, he then assesses the situation as a threat to his wellbeing, at
which point acute distress ensues. This fits well with the proposal on offer: the majority of participants in the experiment appear, at that moment, to lack the resources to withstand the experimenter’s bullying tactics, and, yet, the experiment requires them to do something that they are unwilling to do—something they consider unconscionable. It would therefore make sense that participants would assess the situation as a threat to their wellbeing, and that acute distress would ensue.

There is another, and perhaps more disturbing layer to the experimenter’s contribution to the participants’ distressed state of mind. Once participants were in a state of distress and anxiety, they became more susceptible to intimidation. And the experimenter’s persistence beyond the point at which they became distressed in turn exploited their compromised state, producing epicycles of compliance and ever spiraling levels of distress.

In some cases participants may have even reached a level appropriately characterized as “personal distress,” in the way Merritt, Doris, and Harman define it. Consider, for instance, Bill Menold’s description of how he felt near the end of the experiment, when the learner stopped responding: “at that point I was soaking wet; I was sweating bullets; and, I was starting to laugh almost like a maniac, like hysterically; I’d kind of lost it” (Davis and Perry 2008: 11).

Furthermore, as shock levels rose nearing the final stages of the experiment, some participants began to hesitate less, reverting to what appeared to be survival mode. They seemed to be trying to escape the situation, even if it meant working their way through the levels as quickly as they could (Milgram 1965).
The proposal on offer here, we have seen, accounts not only for the different levels of distress participants experienced throughout the experiment, it also explains the source of their stress, and its progression.

The role the experimenter played in exploiting the participants’ distress perhaps speaks to the aptness of labeling him a “bully” more powerfully than anything else he did. After all, peers, co-workers, and friends sometimes pressure us to do things we do not want to do. Police officers, professors, and parents can be intimidating in various ways. But none of them are bullies simply in virtue of these features alone. The experimenter did not simply pressure and intimidate participants into complying with his unjustified demands, he exploited the fact that they were made more vulnerable by the acute state of distress they were experiencing. Thus, the interpretation offered here gives us a starkly different picture from the standard one used to advertise Milgram’s obedience studies. The participants were not mindlessly obedient to authority; rather they were caught off-guard by the unfamiliar circumstances and the experimenter’s unreasonable behavior—and by the misleading, dubious, and confusing information they received. They were also pressured and intimidated, and experienced severe levels of distress. In short, the experimenter exploited their vulnerabilities and this, I have been arguing, is why participants complied with his unjustified demands against their will.

4. IMPLICATIONS FOR MORAL PSYCHOLOGY

The interpretation of the Milgram experiment defended here has several implications for moral psychology. In the interest of space, however, I focus on the question of whether the results of the Milgram experiment constitute evidence for the claim that moral agents are frail.
The claim that moral agents are frail has gained momentum recently. For instance, Robert Adams remarks, in what appears to be a concession to situationists: “it seems wise to assume that people’s best moral qualities are in some ways and to some degree frail” (Adams 2006: 156; cf. Doris 2010: 137). Of course, whether we should follow Adams (and situationists) on this matter depends on what he means by “frail.” One might, for instance, want to say that Milgram’s participants were frail because they possessed vulnerabilities that the experimenter was able to exploit, and which then lead to them acting in morally impermissible ways (or would have had the shocks been real).

Two problems arise for this suggestion. First, being frail is not the same as possessing vulnerabilities. To see this consider the following analogy: The M1 Abrams tank is an excellent tank. It would not make sense to say that this tank is frail. Nevertheless, it does possess certain vulnerabilities. In particular, if struck with armor piercing shells on the top, flank, or rear of the tank, it could be disabled or destroyed. When something is frail, by contrast, this tends to be because it has many and various vulnerabilities, or an inherent or constitutional flaw of some sort, all of which are difficult (or impossible) to eradicate. A moth’s wings are frail: they can be easily damaged or broken in a number of ways, and there is nothing the unfortunate moth can do about this fact.

Milgram’s participants are not like this. People are vulnerable to being bullied in part because they never learn to develop ways to respond to the tactics of bullies. Often the reason people never learn these strategies is because they are lucky enough not to have encountered many bullies. Sometimes a bad experience with one is all it takes to avoid ever being bullied again. Whatever methods one uses, it would be implausible to say that agents generally do not
have what it takes to stand up to them.\textsuperscript{13} Accordingly, the Milgram experiment is no evidence of the frailty of moral agents.

5. CONCLUSION

When Milgram drew the comparison between the Holocaust and the obedience experiments, he directed our attention toward what participants did, not why they did it. He had us focus on what appeared to be appalling obedience on the part of the subjects of the experiment rather than the experimenter’s unjustified demands. We can be confident that several complex factors ranging over a period of many years came together to convert ordinary men into moral monsters. But Milgram, and others, would have us believe that subtle forces accomplished similar results in the space of an hour-long experiment. I have argued that a very different picture forms once we look at the human face of the obedience experiments. In particular, the puzzle with which we began—namely, how to explain the mismatch between the participants actions during the experiment and their actions in everyday life—is much less mystifying once we learn that participants were made to feel confused, disoriented, pressured, intimidated, and distressed, and that the experimenter achieved widespread compliance by exploiting their vulnerabilities.

What we learn from Milgram’s fascinating series of experiments is, I think, that moral improvement is partly a matter of becoming more familiar with our vulnerabilities. By doing so,

\textsuperscript{13} For similar reasons, I disagree with Neera Bahdwar’s suggestion that the majority of Milgram’s participants have “an unearned sense of [a] lack of self-worth that Aristotle calls pusillanimity” (2009: 283). This would be a kind of inherent frailty that would be difficult to overcome. Nevertheless, if what I have argued above is correct, then Bahdwar’s explanation gets things the wrong way around: while it is likely the case that one \textit{effect} of being bullied is that victims tend to display seemingly meek and defenseless behavior (resembling that of the pusillanimous person), this should be distinguished from the claim that they antecedently possessed the trait of being pusillanimous, and that this in turn played a role in \textit{causing} them to be behave in meek and defenseless ways, which would ultimately explain their compliance. I have provided reasons for thinking that the latter order of explanation is incorrect.
we prepare ourselves for, and fortify ourselves against, the vicissitudes and unexpected dangers of human social interactions.
Bill Menold—a participant in one of Milgram’s original obedience experiments—belongs to a notorious group: he was one of the subjects who completed the experiment by issuing the highest voltage on the “shock generator,” 450 volts, to an unwilling recipient. (He was among the 26 of 40 participants, i.e., 65%, who did the same). In an interview conducted more than forty years later, Bill’s partner, Barbara, was asked about her initial reaction to his involvement in the experiment. She replied in jest: “he courted me by showing me the torture tape” (Davis and Perry 2008: 16). Barbara, it seems, continues to have confidence in Bill’s good character, despite his actions during the experiment. Is Barbara making a mistake? It would be surprising if she were, for Barbara is better placed than anyone to assess Bill’s character. Nevertheless, Peter Vranas makes a plausible case that she is mistaken. On his view, Bill’s actions, along with those of the other compliant participants, lend inductive support to the thesis that most people are fragmented. And fragmentation, Vranas argues, entails indeterminacy. This is the conclusion, he tells us, of a sound argument: the indeterminacy paradox. In all likelihood, then, Bill is, along with most of us, “neither good nor bad nor intermediate” (2005: 1).

In what follows, I argue that despite its initially gripping quality, Vranas’s argument nonetheless falters at its first step. The Milgram experiment (among others) fails to support the thesis that most people are fragmented. The first part of the paper contains my argument against the fragmentation thesis. A crucial part of Vranas’s defense depends on his claim that Milgram’s participants behaved deplorably. Vranas understands a deplorable action as one in which the person is seriously blameworthy for having performed it. I argue that blameworthiness admits of degrees, and that Milgram’s participants were less than seriously blameworthy for their actions.
In the second part of the paper, I argue for my own surprising conclusion: the Milgram experiment supports the thesis that most people are good—not indeterminate; not bad; not even somewhere in between. Accordingly, commonsense character evaluations (e.g., Barbara’s instincts about Bill) may turn out to have some credence after all.

1. THE INDETERMINACY PARADOX

   How would we evaluate someone’s character, Vranas asks, if we learned that the person in question behaved admirably in many situations and deplorably in many others? Someone who would behave deplorably in many situations would not be good, but the same person who would behave admirably in many situations would also not be bad. Notice, too, that such a person is also not intermediate (between good and bad) in the way someone might be if they were only to perform moderate transgressions on many occasions and moderately good deeds on many others. The person we have been describing—someone who is neither good, nor bad, nor intermediate—would be indeterminate. No particular character evaluation would appropriately apply to this individual.

   Of course, it is one thing to admit the possibility of indeterminacy; it is a whole other matter to argue that actual people instantiate this property. Vranas argues, however, that social psychological experiments provide us with evidence in support of the thesis that not just a few, but most people are indeterminate. More specifically, he contends that these experiments support the first two premises in the following argument:

   1. There are many [i.e., in an open list of counterfactual] situations in each of which most people would behave deplorably.
2. There are many [i.e., in an open list of counterfactual] situations in each of which most people would behave admirably.

3. Most people (would) behave deplorably in many (i.e., in an open list of counterfactual) situations and admirably in many other situations.

4. Most people are fragmented. [Follows from 3 by definition.]

5. Fragmentation entails indeterminacy.

6. Therefore, most people are indeterminate (Vranas 2005: 4, 2).

In support of the first premise that most people would behave deplorably, Vranas appeals, mainly, to two social psychological experiments: the first is one of Milgram’s obedience experiments (experiment 5, which Milgram calls the “new base line” condition 1974: 60)), and the second is Philip Zimbardo’s Stanford Prison Experiment. In this connection, Vranas says:

   Clearly, the existence of two situations (Milgram and Zimbardo) in which most people would behave deplorably does not (deductively entail the existence of an open list of such situations. Nevertheless, I think that my argument has high inductive strength, because it suggests that with more ingenuity and less ethical scrupulosity (combined with decreased restrictions from Institutional Review Boards for the Protection of Human Subjects) social psychologists could extend indefinitely the list of such situations.

Accordingly, Vranas suggests that these two experiments provide strong inductive support to the first premise of the indeterminacy paradox.

   In support of the second premise that most people would behave admirably in many situations, Vranas appeals to a different set of experiments. In the first of these, all but one
among the dozens who participated came to the assistance of a technician (actually a
confederate) who, from the unsuspecting participants’ vantage point, appeared to have just been
electrocuted, and would remain in danger if not safely untangled from the wires and electrical
equipment surrounding him. Vranas emphasizes that 71% of these subjects directly intervened
putting themselves at risk of electrocution in order to help the technician (Vranas 2005: 13; Clark
and Word 1974).

In the second experiment Vranas mentions, unwitting bystanders were approached by a
stranger (actually a confederate) who asked them to watch after his belongings until he returned.
Participants agreed, and several minutes later another person (also a confederate) picked up one
of the stranger’s items and started walking away with it. All of those who participated (10 in a
beach setting and 8 in a restaurant setting) ran after the thief (Moriarty 1975). Again Vranas
highlights the fact that “they risked physical confrontation with the thief,” performing an
exemplary action that went above and beyond the call of duty.

I happen to agree with Vranas that these experiments support the thesis the most people
would behave admirably in an open list of situations. In fact, the considerations he outlines in
this connection help lend empirical support to the thesis I defend in the final section of the paper.

Relevant to the next couple sections, however, is the status of his first premise: that most people
(would) behave *deplorably* in an open list of situations. In the third section, I argue that Vranas
lacks sufficient evidence for this premise. The next section, however, is dedicated to recounting
the details of the experiments, and how Vranas sees them as supporting his argument.
2. THE FRAGMENTATION THESIS

Vranas contends that two experiments—Milgram’s baseline condition and Zimbardo’s Stanford Prison Experiment—lend strong inductive support to his claim that most people would behave deplorably in an open list of situations. This claim is indispensable to his defense of the fragmentation thesis because a person is fragmented, according to Vranas, just in case that person would behave deplorably in an open list of situations and would behave admirably in another such open list (2005: 3–4). Without the fragmentation thesis, Vranas cannot establish his conclusion that most people are indeterminate. I turn now to discussing the details of these experiments.

2.1 THE STANFORD PRISON EXPERIMENT

In the summer of 1971, 21 young men (mostly attending universities outside of Stanford, but living in the area for the summer) participated in a simulated prison experiment, staged in the basement of Stanford University’s psychology building.\(^\text{14}\) Selected participants were initially among a pool of 75 candidates who had responded to an advertisement: they would earn 15 dollars per day for (what was intended to be) a two-week long study. 24 respondents were initially selected on the basis of an evaluation of their background, mental stability, and the results of tests indicating that, among the respondents, they were the least inclined to anti-social behavior (21 of 24 went on to participate) (Haney et al. 1973: 73).

Each participant was randomly assigned the role of “prisoner” or “guard.” “Guards” were informed of their roles, and told to attend an orientation the day before the experiment was to

\(^{14}\) The description that follows relies heavily on Haney et al. 1973 and Zimbardo 2007.
begin. “Prisoners” did not have it so easy. After being told to be at their house on a specific day, these young men underwent the following ordeal: with the assistance of the Palo Alto police department, (actual) officers arrived at their homes, announced the charge of burglary, read them their miranda rights, searched, cuffed, and drove them away in the back of a police vehicle, sometimes amidst unwitting family members and prying neighbors (Zimbardo 2007: 29). It was upon arrival in the “prison facility” that the truly degrading treatment began. In front of the “guards” and other “prison staff,” prisoners were “stripped, sprayed with a delousing preparation (a deodorant spray) and made to stand alone naked for a while in the cell yard” (Haney et al. 1973: 76). They were assigned numbers, which appeared on the front and back of the “loosely fitting muslin smocks” or “dresses” they were required to wear. No undergarments were provided. It is important to note that this degrading treatment was built into the experimental design. Nevertheless, it was during this process that some of the guards began to advance their own forms of mistreatment. For instance, Zimbardo reports that a few of them used the opportunity to tease the prisoners, making insulting remarks about their genitals (2007: 40).

Guards were also provided with uniforms: “plain khaki shirts and trousers, a whistle, a police night stick (wooden batons) and reflecting sunglasses which made eye contact impossible” (Haney et al. 1973: 75). They were given very little training or instruction, but were told to “maintain the reasonable degree of order within the prison necessary for its effective functioning,” without the use of physical punishment or aggression (Haney et al. 1973: 74–75).

Despite the lack of instruction, I think it is fair to say that the guards’ sense of how to play their role was shaped by two primary factors. First, they were likely influenced by the portrayal of guards as depicted in the movies, for this would have been their only experience
with what prison-life was like (cf. Zimbardo 2007: 52). Second, after witnessing the initial humiliation of the prisoners during the “induction procedure” (described above), these young men likely drew the conclusion that the type of guard the experimenters wanted them to play was the oppressive kind; one who abuses his power (cf. Banuazizi and Movahedi 1975). In this way, the results of the experiment appear to be self-fulfilling, especially because several of the guards thought that the experiment and its importance hinged on their performance, and so they played their roles with deliberate fervor.

Irrespective of what influenced the guards, it does not change the fact that some of them behaved in despicable ways (while others stood by and allowed it). Vranas provides the following list of examples: Some guards would humiliate the prisoners by “having them enact rituals with a sexual, homophobic character” (Zimbardo and White 1972: 75). They also made them engage in degrading and unsanitary acts by having them “clean out the toilets with their bare hands” (Faber 1971: 83). And, in general, they made their environment extremely uncomfortable in a number of ways including making them do push-ups and refusing permission to have the buckets they had defecated in cleaned out (Zimbardo et al. 1973: 44, 39).

Although each of the guards (roughly 10) engaged in these acts to a greater or lesser extent, Vranas, following Zimbardo, says that “every guard “behaved at one time or other in abusive, dehumanizing ways’” (Vranas 2005: 12; Zimbardo 1975: 45). On these grounds, Vranas contends that most of Zimbardo’s guards behaved deplorably. He also argues that most people would behave as these guards did given that the screening process at the outset of the experiment revealed the young men to be normal, upstanding citizens. It is for these two reasons, according
to Vranas, that the Stanford Prison Experiment lends inductive support to the claim that most people would behave deplorably in an open list of situations.

I discuss my reservations with Vranas’s claims in the third section. Anticipating somewhat, I will argue that the experiment is severely flawed on methodological grounds, and, further, that even if we set aside these problems, the experiment still fails to support the fragmentation thesis because, arguably, most of the guards did not behave deplorably.

2.2 THE MILGRAM EXPERIMENT

In the early 1960’s, Stanley Milgram conducted a series of experiments designed to illuminate the phenomenon of obedience. The basic structure of each experiment was the same.15 Participants responded to an advertisement requesting volunteers—factory workers, businessmen, construction workers, clerks, etc.—for “a study on memory and learning” (1963: 372). When they arrived at the designated location, however, the volunteers learned something more about the nature of the experiment: the goal of the study was specifically to test the effects of punishment, via electric shocks, on memory and learning. Subsequently, each participant was assigned a role—either “teacher” or “learner.” The process of selecting these roles appeared random to the participant, but was, in fact, rigged—the confederate always played the learner. The teacher and learner were then instructed to take up their respective positions. First, the teacher observed as the learner was strapped into a chair and attached with electrodes, at which point the “experimenter” (also a confederate) brought the teacher around to a different room and positioned him in front of the “shock generator.” Here the experimenter gave the teacher an

---

15 The description that follows relies heavily on the discussions in Milgram 1963: 371–376 and Milgram 1974: 13–43).
actual shock of 45 volts (in fact from a battery located behind the shock generator). The function of the shock was to fortify the participant’s belief in the authenticity of the experiment.

The teacher’s instructions were to issue a shock to the learner every time he failed to remember the correct word pair. The cleverly designed, although nonfunctioning, shock generator had small levers corresponding to thirty levels of “voltage” ranging from 15 to 450. In the version of the experiment Vranas focuses on (experiment 5, henceforth “the baseline condition”), the learner lets out a yelp beginning at 75 volts and continues to do so through to 135 volts. Then, at the 150-volt mark, the learner lets out another cry and says:

Experimenter! That’s all. Get me out of here. I told you I had heart trouble. My heart’s starting to bother me now. Get me out of here, please. My heart’s starting to bother me. I refuse to go on. Let me out.

The learner is audible to the participants, but muffled through the walls of the partitioned rooms. At this point, most participants begin to ask questions and exhibit visible signs of hesitation and concern. Six of the forty participants resist the experimenter’s demands to continue at this stage of the experiment—more than at any other level (Milgram 1974: 60).

In response to their questions and hesitations, the experimenter was equipped with a series of scripted prompts, some general and others more specific. The general “prods,” as Milgram called them included: Prod 1: “Please continue,” or alternatively, “Please go on.” Prod 2: “The experiment requires that you continue.” Prod 3: “It is absolutely essential that you continue.” Prod 4: “You have no other choice, you must go on” (Milgram 1974: 21). Among the “special prods,” the experimenter might say “While the shocks may be painful, they are not dangerous” (Milgram 1963: 378). Or: “Whether the learner likes it or not, we must go on,
through all the word pairs” (Milgram 1974: 73; cf. 22). The experimenter’s delivery was not mechanical and it did not have the feel of being scripted. Instead, the prompts blended seamlessly with the various responses of the participants.

Ostensibly, the experimenter would terminate the session if the participant resisted after the four sequence of prompts had been issued after a given level of voltage. If this condition was not satisfied, the prompts would begin anew at the next level. Nevertheless, a consequence of the experimenter’s attempt at realism and a natural delivery was that some participants received more prompts than others depending on how many specific prompts occurred between the more general ones. This varied because different participants responded in different ways. It is therefore the case that this aspect of the experiment was not properly controlled. The fact remains, however, that 26 of the 40 participants in the baseline condition (i.e., 65%) completed the experiment by administering every level of voltage up to and including 450 volts (Milgram 1974: 35).

Vranas’s claim that Milgram’s baseline condition supports the fragmentation thesis relies on two assumptions:

a. Most participants in the Milgram experiment behaved deplorably.

b. Most people would behave as those participants who behaved deplorably did

(Vranas 2005: 4–5).

Regarding (a), Vranas is quite specific about which participants he thinks behaved deplorably. In this connection, he draws two distinctions. The first is between obedient and fully obedient subjects. And the second distinction concerns suspicious versus nonsuspicious subjects. Obedient subjects, according to Vranas, were those who issued shocks of 225 volts or more (i.e., halfway
up the shock generator), while fully obedient subjects were those who went all the way up to 450 volts (2005: 6). Nonsuspicious subjects, Vranas says, were those who “were certain or at least highly confident that the learner was getting painful shocks,” and the remainder of the subjects were suspicious (2005: 6). His thesis (a) is that nonsuspiciously obedient participants behaved deplorably in the baseline condition. This is, he says, “based on the facts that 80% of the 40 participants were obedient (65% were fully obedient) and that according to a follow-up questionnaire (Milgram 1974: 172) about 80% of the participants were nonsuspicious (so that about 64%—i.e., 80% of 80%—of participants were nonsuspiciously obedient)” (2005: 6). In defense of (b) that most people would behave as Milgram’s participants behaved, Vranas notes that Milgram’s sample was large and diverse, and included people who differed in age, occupation, class, level of education, and gender.¹⁶

Vranas expects his readers to share the intuition that nonsuspiciously obedient participants behaved deplorably. I contend, however, that whether we should follow him here depends, first, on certain details of the experiment, and, second, on whether we should accept his account of deplorable action. These are the two central topics of the next section.

3. AGAINST THE FRAGMENTATION THESIS

In this section—paying particular attention to Milgram’s obedience experiments—I argue that we have no reason to think that most people would behave deplorably in an open list of situations because the evidence Vranas cites on behalf of this thesis in fact fails to support it. If this is right, however, then we also have no reason to think that most people are fragmented, and

¹⁶ In the baseline condition, all participants were men, but one variation was identical to the baseline condition except for the fact that all the participants were women. The results were the same: 64% of women participants were nonsuspiciously obedient.
if fragmentation entails indeterminacy as Vranas suggests, then we have no reason for thinking that most people are indeterminate.

In order to evaluate whether most of Milgram’s participants behaved deplorably, we need to know when an action or behavior possesses the property of being deplorable. Vranas is explicit on this matter. He says: “I call an action (token) *deplorable* when it is seriously blameworthy” (2005: 4). As it stands, this characterization is problematic because it makes a category mistake. Agents, not actions, are blameworthy or praiseworthy. Accordingly, I suggest the following amended account:

*Deplorable Action*: An action is deplorable if and only if the agent who performs that action is seriously blameworthy for having performed it.

Now the central question becomes: Were nonsuspicious participants seriously blameworthy for having issued shocks of 250 volts and higher? First, we need to consider whether the participants were blameworthy at all. Here, again, Vranas is explicit. A person is blameworthy for his action, he says, if and only if the action is “(subjectively) wrong (in the sense of violating one’s—all-things-considered—duty) and lacks an adequate excuse” (2005: 4). On Vranas’s definition, most participants were blameworthy for issuing shocks because although the shocks were fictitious (and so their actions were not actually wrong), they were still subjectively wrong in the sense that it was wrong by their own lights to issue what they took to be real shocks to an unwilling recipient. But did they have an adequate excuse? Again, arguably not, on Vranas’s account. As he says: “no dire consequences threatened disobedient participants” (2005: 8). We can also add that participants were not coerced or physically forced into issuing the shocks. Notice, however, that even if most participants were blameworthy, (as far as Vranas has made clear) this does not yet
entail that they were *seriously* blameworthy for issuing the shocks—a crucial qualification in Vranas’s argument. Unfortunately, in a rare moment of imprecision, Vranas fails to tell us when someone counts as being seriously blameworthy for her action, which makes it difficult to evaluate whether the nonsuspiciously obedient participants were in fact seriously blameworthy for their actions.

In what follows, I discuss—and ultimately dismiss—several reasons for thinking that most of Milgram’s participants were seriously blameworthy. Here, Hilary Bok’s illuminating analysis of the Milgram experiment will prove to be indispensable to my argument.

3.1 MILGRAM’S PARTICIPANTS AND ACTING WITHOUT CHOOSING

In her diagnosis of the Milgram experiment, Bok first directs our attention to a phenomenon she describes as *acting without choosing* (1996: 174). Consider by way of illustration a mundane activity such as making a sandwich. Suppose I opt to make a sandwich rather than making soup. I do not thereby have to choose to get the bread out, then the mayonnaise; then the mustard; then the cheese, and so forth. These actions simply follow as a matter of course given my initial choice. Nevertheless, I am acting when I reach for one of the items that will go into my sandwich. I am, therefore, acting without choosing. In general, this way of going about one’s activities promotes efficient use of one’s mental resources. If we had to choose everything we did, much of these resources (both conscious and unconscious) would be dedicated in many instances to deliberating and making decisions rather than to other, potentially more important, tasks. For instance, instead of choosing each and every element of my sandwich, I might instead think about a philosophical problem that has been bothering me.
In certain contexts, however, acting without choosing can get us into trouble. Bok explains that sometimes “we can and do perform actions we know to be wrong simply because we fail to decide what to do” (1996: 174). This can happen when the agent initially chooses a course of action but then the circumstances change in such a way that continuing to perform that course of action is morally wrong. Bok says, when this happens, the agent is “faced with a decisional conflict: a situation in which she has reasons for performing several different incompatible courses of action” (1996: 178). Bok goes on to suggest—rightly, to my mind—that this is precisely the type of scenario Milgram’s participants faced. As she notes, the experiment they volunteered for made no mention of punishment, but was advertised as “a study on memory and learning” (Bok 1996: 179; Milgram 1974: 15). Thus, prior to even arriving at the lab, participants had chosen a course of action that would have them follow the experimenter’s instructions, and, yet, at the time of choice, they had no idea this would entail doing anything immoral.

In fact, the data suggest that most participants were several stages into the experiment before coming to the conclusion that, were they to proceed, they would be doing something morally problematic. In several variations of the experiment, for instance, compliance drops off sharply at the 150-volt mark, which is when the learner cries out in pain and says he no longer wants to take part in the experiment (Packer 2008: 302; cf. Vranas 2005: 8). This is also the point at which compliant participants began to voice their hesitations and concerns about whether the experiment should continue. Accordingly, I think Bok is right to claim that the participants were, at this point, faced with a decisional conflict between incompatible courses of action, and that the
course of action they had already chosen to take, was one that, if they did not decide to do something else, would result in their doing something morally wrong by their own lights.

Where I think Bok’s otherwise incisive analysis misses the mark is in her characterization of the psychology and phenomenology of decisional conflict. In particular, I think she underestimates how difficult such circumstances can be. Consider her description of compliant participants:

Some subjects were able thus to decide what to do, and to act on their decision. But others seem to have viewed their various reasons for action as isolated, conflicting forces between which they were trapped, and not as considerations which they might evaluate and compare. They were, in a sense, paralyzed by their dilemma: *unwilling* to look beyond the mere existence of a conflict to the possibility that they might resolve it” (1996: 181; emphasis added).

Bok describes the participants as willfully unimaginative with respect to resolving their conflict, as if they had the resources at the time to resolve it, and thus were perfectly able to do so had they tried harder, but, instead opted not to make the requisite effort. This idea is also apparent at the beginning of the article when she says that those who fail to extricate themselves from bad situations of this sort, *allow* themselves “to drift” rather than “taking responsibility” for their actions (1996: 174). To see why this fails to account for the mindset of many of Milgram’s participants, it is necessary to consider some of the details of their reactions and behavior during the experiment.

In particular, I want to argue that Milgram’s participants were not willfully blind to a resolution; their passivity did not reflect a failure to take responsibility for the situation; and,
although they willingly chose a course of action that would have them cooperate with the experimenter, at some point they were unwilling to continue to follow his instructions. I argue further that these considerations bear on the extent to which participants’ were blameworthy.

Experiencing a decisional conflict can be an intensely debilitating affair, for it may be accompanied by a feeling of a loss of control over the situation; clouded thought; inability to deliberate; severe distress, and the like. Rarely is it sufficiently detailed how much Milgram’s participants suffered these effects. In a telling description, someone who had witnessed the experiment reported the following:

I observed a mature and initially poised businessman enter the laboratory smiling and confident. Within 20 minutes he was reduced to a twitching, stuttering wreck, who was rapidly approaching a point of nervous collapse. He constantly pulled on his earlobe, and twisted his hands. At one point he pushed his fist into his forehead and muttered: “Oh God, let’s stop it” (Milgram 1963: 377).

As Milgram has noted, reactions of this sort were not uncommon among the participants (1963: 375). Given this fact, it would seem odd if they were simply neglecting to try to resolve their decisional conflict. On the contrary, it appears that participants began to lose sight of their options as they experienced increasing levels of distress and were subject to more and more intense pressure to comply (generated by the experimenter’s seemingly unrelenting demands to continue). Bill Menold—the participant mentioned in the introduction—tells us: “The thought of quitting... It’s really strange. It never occurred to me to just say, “You know what, I’m walkin’ out of here,” which I could have done” (Davis and Perry 2008: 12). The evidence suggests, therefore, that participants were not willfully blind to a resolution. Rather, they were undergoing
an array of complex emotional reactions to the situation and were thus unable to deliberate or think clearly about how to resolve their conflict in an effective way.

The second point concerns the extent to which participants behaved passively during the experiment. It may be tempting to imagine them mechanically “taking orders,” fulfilling the experimenter’s every demand without question. But this could not be further from the truth. In most cases, once participants hear the learner declare that he no longer wants to participate in the experiment, they take some initiative against the experimenter. Often the process begins with hesitation: participants think that the experimenter is going to call it quits in response to the learner’s protests, yet they find—much to their surprise and dismay—that the experimenter wants them to proceed as instructed in spite of the learner’s complaints (Milgram 1965). In certain instances, the participants’ subsequent attempts to get out of it involved desperate exchanges with the experimenter (Milgram 1974: 73–76). Often, their strategy was to point to obvious reasons for terminating the experiment. For instance, that the learner could be hurt or suffer a heart attack and that he was screaming in pain, and so forth. In this way, they attempted to resolve the decisional conflict in a way that would normally seem appropriate; namely through reasoned argument. But, of course, the experimenter would not listen to reason, and for many of the participants, this left them at a loss for how else to respond.

Another question to consider is whether Milgram’s participants were even willing to issue the shocks. This, of course, depends on what one means by “willing.” Consider, for instance, what Bok has to say in this connection:

Obedient subjects who found shocking the learner morally problematic claimed that their actions were involuntary: that they were “totally helpless and caught up
in a set of circumstances where I just couldn’t deviate and I couldn’t try to help,”
or that “I went on with it, much against my will.” Taken literally, these statements
are obviously false. Presumably, what such subjects meant by the claim that they
acted against their will was that they did not want to shock the learner (1996: 181;
emphasis in original).

Bok advises against taking the participants’ remarks at face value. I disagree for several reasons.
First, the phenomenology of decisional conflict (discussed above) reveals why it might have
seemed to participants, at the time, that they could not have done other than they did (in the
psychological, as opposed to the metaphysical, sense). Second, Bok rightly mentions that
Milgram’s participants did not want to issue the shocks, but what she neglects to add is that
issuing the shocks was also against their all-things-considered evaluative stance (cf. Milgram
1963: 376). Sometimes this is all someone means when he says that he acted against his will.
Furthermore, the strongest reason for thinking we should take participants at their word when
they said (or implied) that they were unwilling to issue the shocks is that the experimenter—
despite his apparently polite demeanor—used bullying tactics in his attempts to secure the
participants’ compliance. I have argued at length for this thesis elsewhere, so here I will confine
my remarks to a brief rehearsal of some of the central considerations in its favor.

When I say that the experimenter “bullied” participants, the type of bully I have in mind
is one who, without physical threats, takes advantage of a person’s vulnerabilities in order to get
that person to do his bidding. In what ways were Milgram’s participants vulnerable? In the first
place, they were emotionally vulnerable. We know that the experimental design ensured that
participants faced a decisional conflict. They had already made the choice to cooperate with the
experimenter prior to learning that the course of action they had chosen would involve doing something morally wrong. And I have argued that extricating oneself from a decisional conflict is emotionally challenging in its own right.

The experimenter also played a central role in contributing to the participants’ emotional vulnerability. They naturally expected him to be competent, intelligent, and respectful. Yet he defied their expectations at every turn. They found themselves articulating obvious reasons for ending the experiment to someone who supposedly knew better and, in any case, should have respected their wishes—not to mention the pleas of the helpless learner.

We also know that a consequence of the experimenter’s refusal to heed their requests to discontinue the experiment was that participants became increasingly distressed (Milgram 1974: 42). Pertinent details of this kind are almost universally omitted from descriptions of the experiment. For instance, in his original publication of the study, Milgram tells us that 3 (of the 40) participants experienced “full-blown seizures” (1963: 375). This could hardly be a coincidental effect, and, yet, Milgram neglects to mention it in his 1974 monograph. It raises the question of what other debilitating effects the experiment had on the participants that were never reported. Another example, comes from the interview with Bill Menold, in which he describes how he felt near the end of the experiment: “at that point I was soaking wet; I was sweating bullets; and, I was starting to laugh almost like a maniac, like hysterically; I’d kind of lost it” (Davis and Perry 2008: 11). Despite their emotionally compromised condition, the experimenter continued to press on with his demands—a paradigmatic illustration of a certain kind of bullying behavior.
It was also the case—or so I want to argue—that participants were epistemically vulnerable in various respects. Recent archival work on the obedience experiments reveals that Milgram purposefully instituted several ambiguities into the experiment. For instance, in his initial plans, the end of the shock generator read “LETHAL,” but, as we know, this was replaced with the far more ambiguous, “XXX” (Russell 2011: 149; cf. Milgram 1974: 28). This was part of Milgram’s campaign to achieve just the right threshold between “tension-relieving” factors and “binding” factors (Milgram 1974: 148–152; 157–162). Milgram was highly conscious of the fact that the obedience experiments would have far less impact if everyone or no one complied with the experimenter’s demands (which is precisely what happened in other variations of the experiment).

The upshot of this data is that in the baseline condition, participants were both emotionally and epistemically vulnerable, and that the experimenter took advantage of this fact by pressuring them to comply even beyond the point at which they became confused, disoriented, and severely distressed. His tactics were arguably not coercive, but they were at least tantamount to being just short of coercion. For this reason, I think that once participants came to the conclusion that issuing the shocks violated one of their own moral prohibitions, they were, at that point forward, unwilling to continue issuing them, but were bullied into doing it anyway.

The discussion above is relevant to the issue at hand for the following reason: had most of Milgram’s participants been willfully blind to a resolution; had they not tried to seek a resolution; and had they issued the shocks willingly, then we might want to say, along with Vranas, that they were seriously blameworthy for having issued the shocks (beyond 250 volts). I have argued, however, that most participants fail to satisfy the conditions specified in the
antecedent of this conditional. Thus, they were not seriously blameworthy, *at least on these grounds*. I now turn to arguing more directly for the thesis that Milgram’s participants were less than seriously blameworthy for their actions. In particular, my aim is to defend the claim that when an agent faces *extenuating circumstances*, the degree to which she is blameworthy for her action is either significantly reduced or eliminated altogether. I argue further that the circumstances Milgram’s participants’ faced were extenuating. My argument relies as much as possible on uncontroversial assumptions that are compatible with Vranas’s account of blameworthiness.

3.2 LESS THAN SERIOUSLY BLAMEWORTHY

The first step in the argument will be to use two examples from John Martin Fischer and Neil Tognazzini’s recent paper to illustrate, what I claim are, two sufficient conditions for when a circumstance counts as extenuating. The first example they propose is a familiar enough type of case: a mother has a choice between saving her own child from drowning or saving five other children from drowning, but she cannot do both, and in the end the mother opts to save her own child (2011: 388–389).

As Fischer and Tognazzini note, plausibly the woman’s action in this case is morally wrong, according to standard moral theories, but “such a decision,” they say, “is surely understandable” (2011: 389). It is understandable, they add, “because of the enormous difficulty involved in doing the right thing” (2011: 389). I think Fischer and Tognazzini are right on the mark here. Nevertheless, they neglect to state a further consequence of this line of reasoning; namely that part of what makes the circumstance and her decision so difficult is that she cares so
much for her child. Importantly, it is because of this fact that her action is understandable. If, by contrast, what made doing the right thing difficult for the mother was the fact that her child was a famous actor, and she had tremendous financial incentives to make sure he survived, then the understandability of her action would simply disappear. With this consideration in mind, I suggest the following sufficient condition for when a circumstance counts as extenuating: when it is enormously difficult for the agent to do the morally right thing in that circumstance, and the agent does not possess morally problematic attitudes or motives that render the circumstances difficult.

Fischer and Tognazzini consider a further case, originally articulated by Michael Zimmerman. A man witnesses a car accident, and having seen too many television programs where cars readily explode upon impact, he thinks the vehicle could go up in flames at any moment, so he drags the accident victim from the wreckage, paralyzing him for life (Fischer and Tognazzini 2011: 389; Zimmerman 1988: 41). Fischer and Tognazzini agree with Zimmerman that the would-be rescuer’s action was morally wrong. Working on this assumption, the feature that makes this a case of extenuating circumstances is, in my view, that the man did not have access to certain morally relevant facts; most pertinently, not to move someone who may have suffered a spinal injury. Knowing this fact turns out to be crucial for avoiding moral error in this case. We can stipulate, further, that he was not culpably ignorant for failing to know this fact. Accordingly, another sufficient condition for when a circumstance counts as extenuating is as follows: when the agent fails to have morally relevant evidence about the circumstance she faces,

---

17 It is not a coincidence that they neglect to mention this consequence, for it is a feature of their physiognomy that there exists an unbridgeable gulf between aretaic appraisals of the agent and the agent’s accountability, but this consequence reveals that a feature of the mother’s aretaic structure, i.e., her caring for her child, is after all relevant to her accountability for the action, i.e., to whether or not we are (in their language) justified in deploying the reactive attitudes toward her.
and she is not culpably ignorant for failing to acquire the relevant evidence (or knowledge about) the fact(s) in question.

Fischer and Tognazzini’s examples also illustrate a possible connection between extenuating circumstances and the agent’s blameworthiness for her action in those circumstances. In particular, they may even excuse the agent from blameworthiness. Fischer and Tognazzini conclude something along these lines with respect to the cases described above. Relevant here, however, is whether extenuating circumstances ever mitigate an agent’s blameworthiness for an action. The examples suggest two ways in which this might be the case. The first of these concerns the degree of difficulty: all else being equal, the more difficult it is for the agent to do the right thing (absent morally problematic attitudes or motives), the less blameworthy the agent is for performing the wrong act. Call this the “degree of difficulty condition.” In the second instance, which we might call the “epistemic deficiency condition,” the less evidence an agent has about the morally relevant facts of her circumstance, the less blameworthy the agent is for performing the wrong action in that circumstance (so long as the wrong action is a consequence of this epistemic deficiency and the agent is not culpable for failing to acquire the relevant evidence).

If this is right, then the extent to which Milgram’s participants were blameworthy for issuing the shocks should be significantly reduced on the grounds that they met both the degree of difficulty condition and the epistemic deficiency condition. To see this, consider some of the points previously enumerated. First, we can be confident that doing the right thing was incredibly difficult in the circumstances participants faced. This is illustrated by the simple fact that most participants failed to do the right thing despite the fact that their desires and their all-things-
considered evaluative stance favored doing the right thing. The circumstance was also difficult for the reasons mentioned earlier: they were confronted with a decisional conflict; undergoing intense emotional responses; and targeted with the experimenter’s bullying tactics. Second, regarding the epistemic deficiency condition, participants were confronted with ambiguous and conflicting evidence regarding the morally relevant facts of the case. To give yet another example, the experimenter insisted that the shocks were “painful, but not dangerous” (Milgram 1963: 378), but near the end of the shock board, it read: “danger: severe shock” (Milgram 1974: 28). These considerations, and the ones previously articulated, suggest that most of Milgram’s participants did indeed face extenuating circumstances, and this, in turn, should significantly reduce the extent to which they were blameworthy for issuing the shocks—or so I have been arguing.

Returning finally to Vranas’s argument. Recall that the Milgram experiment purports to lend inductive evidence to the claim that most people would behave deplorably in an open list of situations. And, an action is deplorable if and only if the agent who performs that action is seriously blameworthy for having performed it. Vranas neglects to tell us when someone counts as being seriously blameworthy for performing an action. Rather, he tells us that someone is blameworthy for an action just in case “it is (subjectively) wrong and lacks an adequate excuse” (2005: 8). According to this definition, most of Milgram’s participants were blameworthy, but this is compatible with their being less than seriously blameworthy. I have provided reasons for thinking that this latter claim is precisely the case. If my arguments are on the right track, then Milgram’s nonsuspiciously compliant participants do not lend inductive support to the claim that most people would behave deplorably in an open list of situations
(because they did not act in ways that were seriously blameworthy), and so do not lend support to the thesis that most people are fragmented, and, therefore, do not support Vranas’s conclusion that most people are indeterminate.

3.3 PROBLEMS WITH THE STANFORD PRISON EXPERIMENT

Unlike Milgram’s careful (albeit exploratory) methods, Philip Zimbardo’s Stanford Prison Experiment (SPE) faces several objections on methodological grounds. I argue that, for these reasons, we should not put too much stock in it—let alone accept any unpalatable (paradoxical) conclusions which ultimately rest on it.

The list of methodological problems with the SPE is rather long. Consider, first, the sample size. Only 21 participants eventually took part, and only half of these played guards. Suppose, for the sake of argument, that these 10 or so guards behaved deplorably, Vranas would have us believe that this supports the claim that most (i.e., billions of) people would behave deplorably in a similar circumstance. The representativeness of the sample is also highly problematic. These were all men of a certain age (i.e., 17-30), and overwhelmingly of a similar ethnicity and socio-economic background (i.e., middle class Caucasians) (Haney et al. 1973: 73).

Third, the experiment has never been officially replicated. Arguably, these are the three most basic methodological requirements any scientifically respectable study must meet, and yet it failed all of them. By contrast, the obedience experiments consisted of over 19 experimental

---

18 In this connection, Vranas says the following: “although we lack replication of SPE, a precursor of SPE was carried out in the spring of 1971 by a group of (both male and female) undergraduates who had been assigned in one of Zimbardo’s courses the project of studying prison life (and who, incidentally, “belonged to a dormitory house plan which was dedicated to nonviolence”). The results were apparently similar to those of SPE: “the mock guards dehumanized the mock prisoners in a variety of ways” (Zimbardo 1975: 37; Vranas 2005: 12). It is difficult to evaluate whether this helps Vranas’s case. The sample is still minuscule and not fully representative. Furthermore, we would want to know more of the details of the experiment and what constituted “dehumanizing” behavior, according to Zimbardo.
controls, contained a large sample size (i.e., usually 40 participants per experiment, but never less than 20); the participants varied in age, gender, ethnicity, education level, and socio-economic status; and, lastly it was replicated many times over in experiments conducted all over the world.\textsuperscript{19}

Vranas addresses some of the methodological problems with the SPE. He says that although the sample was not representative, the extensive testing the men underwent at the outset provides some evidence that they were like any other ordinary, morally upstanding citizens. The problem with this reply is that it conflicts with other things Vranas says about the participants. Following Zimbardo, Vranas says: “\textit{many} guards did not harass prisoners unwillingly: they reported “being delighted in the new-found power and control they exercised and sorry to see it relinquished at the end of the study”” (Vranas 2005: 11; Zimbardo et al. 1973: 49; emphasis added). Once again, citing Zimbardo, he adds: ““most of the worst prisoner treatment came on night shifts and other occasions when the guards thought they could avoid surveillance and interference of the research team”” (Vranas 2005: 11; Haney et al. 1973: 92). If these descriptions of the SPE participants is accurate, then it spells trouble for Vranas’s argument. To see why, it will be helpful to return to Milgram, and, in particular, to the results of two of his experiments.

Milgram’s so-called “free choice” condition was similar to the baseline condition in most respects except that participants were allowed to choose the shock level. The result was that only 2 of 40 participants issued shocks above 150 volts (Milgram 1974: 61). We also know that in another condition, where the experimenter was absent—and so subjects thought they were not

\textsuperscript{19} Vranas (2005: 9) provides a helpful list of citations of the various locations where the experiment has been replicated.
under any surveillance—“several subjects administered lower shocks than were required and never informed the experimenter of their deviation from the correct procedure” (Milgram 1974: 62). Milgram, and others, have suggested (rightly in my view) that these experiments reveal that ordinary people tend to have certain morally respectable values, such as the prohibition against harming an innocent person (cf. Merritt, Doris, and Harman 2010: 364). For instance, Milgram says:

> It is clear from the remarks and outward behavior of many participants that in punishing the victim they are often acting against their own values. Subjects often expressed deep disapproval of shocking a man in the face of his objections, and others denounced it as stupid and senseless (1963: 376).

The evidence here, together with the reasons for thinking Milgram used a representative sample, suggest that most ordinary people when left to their own devices, do not behave aggressively (Milgram 1974: 70–71), and do not violate the prohibition against harming innocent people, even when they have the power to do so, and they are confident they can get away with it.

By contrast, Zimbardo’s description of the guards suggests that his participants behaved aggressively and willingly (and happily) engaged in actions which were (psychologically) harmful to the prisoners without much (or any) prompting. The discrepancy between Milgram and Zimbardo’s results can be explained in three distinct ways: (1) Zimbardo’s sample of participants is not representative of the general population. (2) The SPE lends inductive evidence to the thesis that most people are *bad*, not indeterminate. (3) The description Zimbardo gives of the participants (the ones Vranas cites above) do not accurately portray how most of the guards
acted and behaved. Each explanation poses a problem for Vranas’s argument, and the list appears to be exhaustive.

Consider (1). Supposing that Zimbardo’s description of the guards is correct, and given that Milgram’s data is based on a representative sample, this suggests that Zimbardo’s sample is not representative. After all, ordinary people, as we have seen tend to have morally respectable values, attitudes, and dispositions, which make them averse and unwilling to perform harmful actions. Yet, Zimbardo’s description of the guards suggests just the opposite, so his sample must not be representative of the population at large.

Alternatively, consider option (2). Again, if we take Zimbardo’s description of the guards for granted, and supposing, for the sake of argument, that by a mysterious coincidence Zimbardo’s sample was representative, while Milgram’s was not, Vranas’s argument would still be in trouble. Here’s why: someone who willingly, and even happily, engages in deplorable actions is not indeterminate, but bad. Thus, if Zimbardo’s guards were indeed representative of the population at large, and they engaged in deplorable actions with sadistic fervor with little to no prompting, then it would not support the conclusion of Vranas’s indeterminacy paradox. Rather, it would support the conclusion that most people are bad.

Perhaps, then, Zimbardo’s portrayal of the guards as aggressive and sadistic is not entirely accurate. In my view, this is the option most likely to be true. And, in places, Vranas appears to agree; following Zimbardo he says:

It is true “about one-third of the guards were “so consistently hostile and degrading as to be described sadistic. They appeared to take pleasure in the prisoners’ suffering” (Zimbardo 1975: 46;). But other guards were “tough but
fair” (‘played by the rules’), ...while a few [or “several” Zimbardo 1973a: 154; Zimbardo & White 1972: 70] “were passive and rarely instigated coercive control” (Haney et al. 1973: 81): “they occasionally did little favors for the prisoners, were reluctant to punish them, and avoided situations where prisoners were being harassed” (Zimbardo et al. 1973: 49; Vranas 2005: 11–12; brackets in original).

The passage leaves much to be desired. Given how few guards there were (roughly 10), a little more precision on Zimbardo’s part would have been helpful. Instead, he uses vague quantifier terms, such “many,” “most,” and “several,” which means that at any given time we cannot be sure if the actions and behaviors he describes were performed by less than a majority; a bare majority; or a super majority. But knowing which of these is actually the case is crucial to evaluating Vranas’s argument. Setting this aside, however, Zimbardo does seem to be suggesting in this passage that only one-third of the guards (not “many” as was previously suggested) behaved deplorably, in which case the SPE would not support the thesis that most people would behave deplorably in a open list of situations.

Vranas has a ready reply to this objection: “every guard,” he says quoting Zimbardo, “behaved at one time or other in abusive, dehumanizing ways” (Zimbardo 1975: 45), and that “even those ‘good’ guards...respected the implicit norm of never contradicting or even interfering with an action of a more hostile guard on their shift” (Haney et al. 1973: 94)” (Vranas 2005: 12).

Again, regarding the first point, the devil is in the details. In order to evaluate Vranas’s reply, it would be necessary to know the precise nature of these supposed abuses committed by the “good” guards, but such details are not provided.
Another problem, at least on the face of it, is that it is sometimes difficult to distinguish between the dehumanizing activities of the guards from the prison environment itself. Prisoners, unlike guards, were not allowed to leave. Instead, they had to serve their prison sentence in humiliating uniforms, and confined spaces, with no privacy or basic accommodations. For example, they were given “three supervised toilet visits” a day (Haney et al. 1973: 76; emphasis added). The experiment itself was arguably deplorable, and participants had some complicity in this, but I would contend that it is the experimenters who designed and executed it that were seriously blameworthy for doing so. By contrast, most of the young men recruited as guards—and who we have reason to believe tried to avoid harassing the prisoners—were less blameworthy for the actions they performed in the wake of an already corrosive experimental design. If this is right, and most guards were less than seriously blameworthy, then on Vranas’s account, their actions were not deplorable.

Lastly, Vranas’s second remark about the guards raises a further problem: he suggests that omissions not just actions can be deplorable. On this view, the guards who retreated to the guard’s station (Zimbardo 2007: 188) when others began harassing the prisoners were, nonetheless, seriously blameworthy for many deplorable omissions. It is at least controversial whether we should share Vranas’s intuition here. Furthermore, once omissions get included into the fold, Vranas’s case is easy: almost all of us, all of the time have countless deplorable omissions to our name. In fact, I doubt anyone would be left unscathed by Vranas’s argument (not even moral saints). Thus, if one is disqualified from being a good person on the ground that one is responsible, and, thus, seriously blameworthy for an open list of deplorable omissions, then it is hardly a paradoxical conclusion that most people are not good, but indeterminate.
I take it that the real lesson of the Stanford Prison Experiment (hardly one we needed an experiment to teach us) is that when institutions lack the proper mechanisms to prevent abuses of power, they occur. This is evidence that, over time, people are corruptible by unjust institutions. Ordinary people might become fragmented or bad through certain influences, but this does not show that these same people are bad or fragmented. I raise this point again in a different context in the section that follows.

In this section, I argued that the SPE faces extensive methodological objections. I argued, further, that even on the most charitable interpretation of the experiment, it still fails to lend support to the claim the most people (would) behave deplorably in an open list of situations. If this is right, then it also fails to support the fragmentation thesis, and, therefore, the conclusion that most people are indeterminate.

4. OBJECTIONS AND REPLIES

I now turn to addressing two objections to my arguments thus far. The first of these denies that there are degrees of blameworthiness, and suggests instead that to be blameworthy just is to be seriously blameworthy. Accordingly, given that Milgram’s participants performed an action that was subjectively wrong, and lacked a relevant excuse, they were, on this account, seriously blameworthy. Vranas would, of course, be free to stipulate these notions in any way he preferred.

Suppose for the sake of argument that Milgram’s participants did behave deplorably in the baseline condition. Milgram conducted over nineteen other experiments of a similar variety, and often participants did not behave deplorably on Vranas’s definition. Furthermore, recent
archival research reveals that Milgram worked hard to get the details of the experiment *just so* in order to achieve the desired level of compliance (Russell 2011). These considerations mount a case for thinking that most people will behave in the way participants did in the baseline condition only in a highly specific type of circumstance. If this right, however, then the obedience experiments (taken as a set) would not lend inductive support to the claim that there is an *open* list of situations in which most people would behave deplorably. Instead, it would lend inductive evidence to the contrary claim that there is a *closed* list.\(^{20}\)

The second objection I want to consider to my argument runs as follows: the Milgram and Zimbardo scenarios are just two among many in which people (are purported to) behave deplorably. There are a number of other cases, some of which are more “naturalistic,” to use Vranas’s term (2005: 27) that would work just as well in support of the fragmentation thesis.

For obvious reasons, my reply will not address every purported example of people behaving deplorably, but I will address the ones Vranas mentions. I argue, further, that they reveal a pattern that suggests that any experiment or real-life case that Vranas might attempt to use in support of the fragmentation thesis will likely fail on the very same grounds as the cases I am about to discuss.

Vranas describes two more experiments, which are said to have the advantage of being more realistic, and are, therefore, less likely to produce artificial effects owing to suspicious participants. In the first of these (Hofling et al. 1966), an experimenter posing as a physician phoned nurses at two separate hospitals, and instructed them to administer an excessive dose of

---

\(^{20}\) Vranas would face another problem if being blameworthy entailed being seriously blameworthy; namely accommodating the cases raised by Fischer and Tognazzini. Arguably on Vranas’s account, the mother and the would-be rescuer did something (subjectively) wrong and lacked a relevant excuse. I say this because Vranas appears to be taking a hardline on excuses, for he suggests that the only thing that would have excused Milgram’s participants is if they had been threatened for noncompliance. If, on the other hand, Vranas wanted to suggest that the mother and the would-be rescuer did have an excuse, then he would owe us a better account of why Milgram’s participants did not also possess one in light of the considerations I have raised.
an unfamiliar medication to a patient under their care. 21 of 22 nurses were about to comply before being stopped by another experimenter. Had they been allowed to comply, they would have violated official hospital policy against administering medication on the basis of telephone orders. In a second experiment (West, et al. 1975), a person disguised as a government official (actually a confederate) approached criminology majors in an attempt to persuade them that a local advertising firm was alleged to contain illegal records of a plot to defraud the government of 6.8 million taxpayer dollars. Their task was to break into the firm and microfilm the relevant documents. 9 of the 20 participants who were guaranteed immunity from prosecution agreed to do it, whereas only 1 of 20 who were warned they could not be guaranteed immunity agreed.

Vranas’s choice of examples is potentially revealing. If there really is an open list of situations in which most people would behave deplorably, one would expect it to be relatively easy to generate a wide-range of uncontroversial examples, in which people behaved deplorably —especially given the amount of energy social psychologists have dedicated to conducting experiments of this kind. Yet, both experiments suggest that it is not the easy to find cases of people behaving deplorably. For instance, the breaking and entering case, is not an uncontroversial case of deplorable action. To see this, it is necessary to get clear about what Vranas must think the deplorable-making feature of their action is. It cannot be the act of breaking and entering (or the ease with which they agreed to break and enter) because 11 of 20, i.e., a majority of the participants declined; so, on this supposition, the case would fail to illustrate that most people would behave deplorably in an open list of situations. (In fact, if breaking and entering was the deplorable-making feature, this experiment would constitute evidence against the fragmentation thesis.) Instead, the feature Vranas places emphasis on is the
idea that people are more inclined to do morally dubious actions when no negative consequences will come to them as a result. Are the college-aged participants in this experiment really behaving deplorably for failing to do something that could do the taxpayers some good, but is not worth the risk of imprisonment, or disrupting their lives and relationships? I contend that this is certainly not an uncontroversial case of deplorable behavior.

The discussion of these two experiments suggests the natural thought that it is not in fact that easy to find cases in which most ordinary people (would) behave deplorably. One plausible explanation of this fact is that no open list of such situations exists after all.

Another salient feature worth noting in the telephone medication experiment is its striking resemblance to the Milgram paradigm. Much like the participants in the obedience experiments, the nurses were in an epistemically compromised position; they were under pressure to make a decision contrary to the purported expert and authority, and so they inevitably second-guessed their own judgment. As in the Milgram experiment, this would not excuse the nurses from blameworthiness for their professional negligence, but it does call into question whether they were seriously blameworthy for intending to comply; and, so, whether their intended actions were really deplorable.

Vranas raises yet another case, which he explicitly compares to the Milgram experiment. He says: “an important factor in as many as 25% of all airplane accidents is excessive obedience to authority” (2005: 27; Tarnow 2000: 120). Again, one suspects that if there were a wide-range of experiments, in which people behaved deplorably in variety of situations, not just ones that resemble the Milgram paradigm, then Vranas would mention it. In the absence of such evidence, the following hypothesis looks more likely to be true: there is a closed list of highly specific
situations in which most people are prone to moral error—namely, circumstances in which an
agent must do the right thing without all the morally relevant facts, while also under time
constraints or pressure to opt for a course of action in the face of a presumed expert or authority
demanding that they do the wrong thing. If this is right, we do not have reason to think that most
people are indeterminate, but, rather that they are susceptible to moral error in circumstances of a
highly restricted kind. In the remainder of this section, I address a final set of cases Vranas raises
that could, however, call this hypothesis into question.

It is a well-known fact that atrocities of all kinds have regularly taken place, spanning the
globe, as far back as the historical record goes. Vranas reminds us of two such examples: the
actions of the German Reserve Police Battalion 101 in the summer of 1942 and the actions of the
Hutus in Rwanda during a three-month period in 1994.

The mostly middle-aged men of the Reserve Police Battalion, many of whom had never
discharged their weapons in a combat situation, would go on to complete their first of many
murderous assignments, in Józefów, by shooting an estimated 1,500 men, women, and children,
all Jews, at point-blank range in the back of the neck (Browning 1993: 64; 225). In the Rwandan
Genocide, Hutu citizens murdered their Tutsi neighbors (men, women, and children), hacking
them to death with machetes (Gourevitch 1998: 191–194). In the end, the death toll mounted to
an estimated 800,000 people over the course of a few months (Gourevitch 1998: 3).

One of the most disconcerting features of these atrocities is that those committing the
deplorable acts appear to be “ordinary” people. If this is the correct description, then these cases
would support Vranas’s claim that most people (would) behave deplorably in an open list of
situations.
It goes without saying that the actions of those who participated in these atrocities are unquestionably deplorable. The question is whether these people really are ordinary, and what is “ordinary” anyway? One way to get a sense of what ordinary people are like is to rely on experiments. The advantage of the laboratory is that it is possible to select a large and diverse sample of participants, and to control the environment in order to see what they would be willing to do of their own accord during a short time without the worry that the experimental setting has actually changed the participant in a fundamental way. Recall that in the free choice condition this is precisely what Milgram did, and only 2 of 40 participants issued shocks above 150 volts. Furthermore, if my interpretation of the baseline condition is right, participants did not willingly issue shocks beyond 150 volts. Rather, the experimenter had to engage in tactics just short of coercion, i.e., he had to bully them, in order to get participants to comply. Accordingly, evidence from the obedience experiments suggests all of the following: ordinary people are not inherently aggressive; they possess respectable values, attitudes and dispositions, as well as a range of moral emotions which inhibit immoral behavior; and, finally, they are not readily induced to engage in serious wrongdoing. Thus, there is both a normative and statistical sense of ordinary relevant here: all else being equal, most people possess morally respectable values, attitudes, dispositions, and moral emotions that render them unwilling and averse to performing deplorable actions.

Returning now to the perpetrators of the atrocities in Józefów and Rwanda. These were people whose values, attitudes, and dispositions had been shaped and distorted in problematic ways by both longterm and short-term influences. In Germany, there was a history of anti-Semitism and anti-Bolshevism (in many ways the latter amounted to the former because it
targeted Jewish intellectuals). In Rwanda, there had been longterm civil unrest between the Hutu and the Tutsi, fueled over the years by the social, political, and economic successes of the Tutsis.

In the short run, there were triggering events. In Germany, this was the rise of National Socialism, and Hitler coming to power. In Rwanda, it was the death of Hutu president, Habyarimana, in a suspected assassination. Even more immediate were the affects of dehumanizing propaganda and fear mongering. In Germany this was the supposed “partisan” or Bolshevik threat, and in Rwanda, the extremist-run radio (falsely) warned Hutu citizens that it was “kill or be killed” (Gourevitch 1998: 95). In short, these were not ordinary people, but rather individuals whose attitudes, values, and dispositions were already problematic in various ways, and continued to be altered for the worse by malicious influences. In other words, the people who committed these atrocities did (and would) behave deplorably because they were no longer good people, and their values, attitudes, and dispositions reflected this fact.

It is precisely because those who played a part in these atrocities were not ordinary that they fail to lend inductive support to the fragmentation thesis. For, on Vranas’s account, a person is currently fragmented just in case that person, as they currently are, (i.e., with their ordinary and morally respectable attitudes, values, and dispositions) has (or would) behave deplorably in an open list of situations (2005: 4). But the examples given above illustrate that individuals through a fundamental change in their attitudes, values, and dispositions can become the kind of people who would behave deplorably in an open list of situations. This is compatible with the claim that most ordinary people can become fragmented (or bad, or downright evil), but it does not support the claim that most people are fragmented. The latter claim, however, just is the
fragmentation thesis, so the atrocities Vranas mentions do not support his argument.

5. THE GOODNESS OF ORDINARY PEOPLE

In this section, I defend an argument that, if sound, would show that the Milgram experiment (baseline condition) lends inductive evidence to the claim that most people are good. My argument consists of four premises and a conclusion.

*First premise: Most of Milgram’s nonsuspiciously obedient participants were either good, bad, somewhere in between good and bad, or indeterminate. (Estimating 4% pathological participants the total is 60%).*

This premise claims that almost all of the nonsuspiciously obedient participants were moral agents, and, therefore, met the conditions necessary for attributions of moral blameworthiness (and praiseworthiness). It assumes, further, following Gary Watson, that one of these necessary conditions is that the subject be the appropriate target of aretaic appraisals (1996: 266). One way to interpret Watson’s thought is that all moral agents are, necessarily, either good, bad, somewhere in between good and bad, or indeterminate.21 I take this to be an exhaustive list of possible character evaluations (in the broadest terms), and I believe this fits well with Vranas’s understanding.

I will not spend any time defending the account of moral agency the first premise rests on, first, because it is relatively uncontroversial (cf. Fischer and Tognazzini 2011: 382–385), and, second, because Vranas must be assuming something along the same lines; for he thinks that

---

21 Notice that these notions might in turn be gradable: some people are better (more good) than others, while other people are worse (more bad) than others. Accordingly, carving up the aretaic terrain in these broad terms is not as crude as it may appear on the surface.
nonsuspiciously obedient participants are morally blameworthy for their actions, so he, too, is presupposing they are moral agents.

As for the premise itself, and, in particular, my estimate of the percentage of pathological subjects, my evidence comes from the free choice condition. Recall that only 2 out of 40 participants (i.e., 5%) issued shocks of more than 150 volts when they were allowed to choose the level of voltage. Calculating for suspicious subjects (80%), this leaves us with 4%. Thus, the percentage of nonsuspiciously obedient participants who were either good, bad, somewhere in between good and bad, or indeterminate is 60% (i.e., .64 − .04 = .60).

Second premise: Most of the nonsuspiciously obedient, non-pathological participants (60%) were either good or somewhere in between good and bad. (The bare majority of participants in this set is, therefore, 31%).

This premise claims that most (i.e., on a extremely conservative estimate, a bare majority 31%) of the non-pathological, nonsuspiciously obedient participants were either good or intermediate (between good and bad). Defending this premise requires arguing against Vranas’s implicit denial of the claim that a person’s character is grounded in her morally relevant attitudes and dispositions.

Vranas’s view appears be that fragmentation is a brute fact; it is not in virtue of some further fact about a person’s attitudes and dispositions that she is fragmented. I say this because he thinks ordinary people with morally respectable attitudes and dispositions can be fragmented, and, therefore, indeterminate. But, arguably, if a person is fragmented in virtue of her morally relevant mental states, such as her attitudes and dispositions, then these very same mental states
will also be fragmented, and Vranas denies this latter claim. The following example illustrates why Vranas’s implicit assumption seems false.

Consider someone who is actually (as opposed to counterfactually) fragmented. This is someone who actually behaves deplorably in a number of situations and admirably in a number of others. The fictional example of Walter White from the television program “Breaking Bad” should serve my purposes here. Walt is a high school chemistry who is diagnosed with terminal lung cancer. With a baby on the way; a son with Cerebral Palsy; and no savings, he comes to the rather drastic decision to start his own Methamphetamine lab, and to keep his illegal activities a secret from his family. His plan is to find someone to help him distribute the Meth (a former student as it turns out) in order to make large sums of money very fast, so as to provide longterm financial support for his family in his absence.

Walt’s decision, and its execution, very quickly lead him down a path to murder, lies, and severe risks to his own wellbeing, as well as that of his business partner, and, ultimately, his family. At the same time, however, he risks his own life at various junctures to save the lives of both his business partner and the members of his family. Arguably, Walt is fragmented. Notice, however, that his values, attitudes, and dispositions also change in subtle, gradual ways. In particular, they become more contextually sensitive. For instance, he appears to hold a strict prohibition against harming people in his ordinary, “suburban life,” but he does not hold the same prohibition in his “street life.” Some of his moral transgressions do seep into his family life: most notably the increasingly extravagant lies he must tell to maintain his various secrets. But these are often a response to, and follow in the wake of, pronounced changes in his street life. Accordingly, a natural explanation of why Walt is fragmented is because his morally
relevant attitudes and dispositions are fragmented and his morally relevant attitudes and dispositions are fragmented because his life is fragmented. He is living in two worlds.

Vranas owes us a plausible moral psychology, according to which ordinary people with morally respectable attitudes and dispositions are also (at the same time) fragmented. In the absence of such a story, I think it is safe to assume that bad and fragmented people fail to possess (only) these attitudes and dispositions. If this is right, however, then most of Milgram’s nonsuspiciously obedient participants, who we are confident did possess morally respectable attitudes and dispositions, were either good or intermediate (between good and bad) in virtue of these very same attitudes and dispositions.

The third premise: Of the participants who were either good or somewhere between good and bad, most of them were good. (Conservative estimate: \( .71 \times .31 = 22\% \)).

My evidence for this premise comes from the very data Vranas uses to support his thesis that most people (would) behave admirably in an open list of situations. A strong majority (at least 71%, and in other cases 100% of subjects (Clark and Word 1974; Moriarty 1975)) went beyond the call of duty to help someone. Yet, I contend that someone who would behave admirably in a number of situations, and who never behaves deplorably is good—not intermediate. We therefore have reason to think that among the participants who were either good or intermediate, most of them would have behaved as those participants who behaved admirably did, and for this reason they are good rather than intermediate.
The fourth premise: Nonsuspiciously disobedient subjects were good (.36 \times .80 = \text{roughly } 29\%).

The fourth, and last premise says that disobedient participants—at least the ones who were not suspicious of the experimental ruse—were good. (Some of them who were in fact suspicious were likely good as well, but setting those to one side, my highly conservative estimate is roughly 29\%, i.e., .36 (the percentage of disobedient subjects) \times .80 (the number of suspicious participants). The evidence for this claim is fairly transparent: despite the difficulty of the circumstances, these participants managed to stand their ground against the experimenter’s bullying tactics, and to do what was right. This strikes me as a testament to their good character.\textsuperscript{22}

Conclusion: Therefore, the majority of Milgram’s participants in the baseline condition were good. (Conservative estimate: .22 + .29 = 51%).

One possible response to this argument is that, even if it is sound, it would only establish that barely a majority of people are good. True, but the premises are based on the most conservative of estimates. There is some reason to think the numbers are in fact higher. For instance, it is probably the case that many (i.e., a large majority) of the nonsuspiciously obedient participants possessed morally respectable attitudes and dispositions given the results of the free choice condition, where only 2 of 40 participants did not appear to be at all inhibited by their attitudes and dispositions. Additionally, I used the lowest estimate available from Vranas’s data to provide evidence for their goodness over their intermediateness; namely 71\%, but some of the

\textsuperscript{22} This is compatible with obedient participants being good because, on my view, character evaluations are gradable; so disobedient participants are even more good than the relevant obedient participants.
data are as high as 100%. Each of these factors would boost the number of participants who were good substantially above 51%.

I thereby contend that the conclusion of the argument presented here lends inductive support to the claim that most people are good. In my final remarks, I will provide reasons for thinking that this result generalizes to empirical data across the board.

In the course of arguing against the fragmentation thesis, an important pattern emerged. The various experiments and real-life examples Vranas raises all appear to fit the description of one of the two following types of case:

Case 1: Good people with morally respectable attitudes and dispositions will, in certain circumstances, act in ways that violate their own moral prohibitions.

Case 2: People who were perhaps once good, undergo a fundamental change in their values, attitudes, and dispositions; and, so behave deplorably in a number of circumstances as a result. These are not good people (nor are they between good and bad). They are either people who become bad (e.g., perpetrators of atrocities) or they have become indeterminate (e.g., Walter White).

If this is right, however, why has it seemed to some philosophers that social psychological experiments, in particular, have been the key to undermining commonsense character evaluations. I need to explain away the appearance that the copious experiments in social psychology and numerous real-life examples show more than I have suggested.

I contend that social psychologists themselves have played a part in producing certain illusions. They have done so, first, by getting us to focus exclusively on the outcome of these experiments rather than how the experiment was designed in part to produce these outcomes.
Second, they have found clever, and sometimes unexpected ways, to construct the experimental design in order to influence people to act in the kinds of morally problematic ways we would observe had they undergone an equally problematic change in their attitudes and dispositions. In other words, because the participants in these experiments have *not* undergone such changes, those who design the experiments are required to construct them in such way that features of the situation will do the same work that the participants’ morally corrupt attitudes and dispositions would have done, had they possessed them. I argued above, that this requires isolating and targeting our vulnerabilities (both emotional and epistemic) and (ironically) taking advantage of our good nature (i.e., our kindness, our willingness to be cooperative, our shared conceptions of the importance of scientific advancement, our norms of etiquette, and so forth). If, however, the actions of the participants in these experiments owe more to the difficult and unusual circumstances than they do to that person’s character, then by the same reasoning their actions do not reflect their character (or lack of character) in these circumstances. But this suggests that we should not infer the presence (or absence) of character from situations of this sort, and, yet, this is precisely what Vranas (and other like-minded philosophers) have attempted to do.

The real threat to character are the social, political, and institutional influences that, overtime, are capable of corrupting good people. Does this possibility speak to the nature and robustness of our character? Certainly it does, but in a familiar enough way: most people are not moral saints; they are not incorruptible; and they are not immune from moral error in extenuating circumstances. And, yet, the social psychological experiments should be a source of optimism: most people possess the various attitudes and dispositions in virtue of which they are good, and
the circumstances must be quite unique, rare, and difficult for these attitudes and dispositions not to manifest in action.
CHAPTER 3: FAILURES TO MANIFEST OR MANIFEST FAILURES? NEW DIRECTIONS FOR A DISPOSITIONAL THEORY OF CHARACTER TRAITS

Recent years have seen a resurgence of interest in character traits, and the dispositional theory in particular. In this connection, authors appear to be drawing from, and in some cases combining, what I will argue are importantly different traditions. The most familiar (and longest running) of these is found in the writings of the ancient Greek philosophers, especially Aristotle (1985). Another, more contemporary, tradition is found in the psychological literature, most notably in Lee Ross and Richard Nisbett’s (1991) pioneering defense of the thesis known (within philosophical circles) as “situationism.” A third tradition—running parallel to the other two—dates back to Gilbert Ryle (1949), and, more recently, Richard Brandt (1970).

In order to illustrate a few of the dimensions along which these traditions diverge, consider, first, one mark of the Aristotelian conception of character traits. On this view, traits of character are moral virtues or vices, and as such are included in a web of other concepts relating to human excellence and the question of “How to live?” One drawback of this tradition—from the standpoint of a proponent of the dispositional theory of character traits—is that one might want to acknowledge an important explanatory and normative role for character traits without also affirming Aristotle’s substantive views on matters such as: the unity of the virtues, the doctrine of the mean, and incontinence—to give just a few examples.

Now consider Ross and Nisbett’s assumptions about the nature of character traits, in what is, broadly speaking, a critical treatment of the topic. In places, these authors seem to assume that the dispositions associated with various traits are not capable of being penetrated by environmental, cultural, or “situational” influences, and that dispositions—as if they were some kind of innate structure—would fully determine every aspect of a person’s actions and behavior.
regardless of certain features of the circumstances, if, in fact, that person possessed the relevant behavioral disposition (cf. Ross and Nisbett 1991: 4). Relatedly, and along similar lines, they sometimes seem to imply that there is a strict dichotomy between dispositions and situational factors: either the former or the latter explain behavior, not some combination of the two (Ross and Nisbett 1991: 59). Another striking feature of their account is their assumption that whether a person possesses a disposition to behave depends, in part, on the consistent manifestation of that disposition, and, so, on this understanding, we should be able to predict with stunning accuracy a person’s overt behavior (Ross and Nisbett 1991: 2–4).

Lastly—and in stark contrast with the former approach—Ryle argued that we can make sense of character traits in purely dispositional terms, which are “not themselves [either] overt or internal” (1949: 33). He means they are neither internal in the sense of being dependent on some immaterial substance, and, yet, neither are they overt in the sense of reducing to observable behavior. For his part, Brandt explicitly rejects Ross and Nisbett’s second assumption that consistency in overt behavior ought to be built into the dispositional account of character traits.

Accordingly, the extent to which these three traditions are talking about the same phenomenon is, at best, questionable, and, yet, recent discussions of character traits tend to propose accounts that blend and draw from all three traditions in a way that has resulted in much confusion and ambiguity—or so I will argue.

The first step toward greater clarity is to try to understand the challenges faced by proponents of dispositional accounts of character traits in a more theory (or tradition)-neutral way. The problem, as I see it, is not, as some authors have suggested, that dispositional theories
of character traits are subject to empirical falsification (cf. Doris 2002: 15–16). Rather, the problem is the *prima facie* incompatibility of the following three claims:

1. *The dispositional theory of character traits:* A subject possesses a character trait just in case that person possesses a certain set of dispositions to think, feel, and act in certain ways in certain circumstances.

2. *The masking thesis:* It is possible for a person (or object) to possess a disposition that never (or hardly ever) manifests, indeed, the normal case is that they frequently fail to manifest owing to such things as masks (i.e., factors that interfere with the manifestations of dispositions).

3. *The empirical thesis:* There is a strong empirical connection between possessing certain character traits and performing certain actions.

In the first section of the paper, I explain why someone might want to hold each of these claims. Taken together, they appear to be incompatible, so I will call this “the trait-action problem.” In the second section, I critically evaluate one familiar attempt to resolve the trait-action problem. I conclude that it is not successful. In the third, I defend my own resolution.

My primary aim, then, is to articulate the most promising strategy for the proponent of a dispositional theory of character traits, and this crucially involves resolving the trait-action problem. Along the way, however, I also pursue several subsidiary aims. First, I intend to address certain under-discussed, and often implicit, metaphysical presuppositions found in the literature on character traits, especially those whose misunderstandings have posed problems for the viability of dispositional theories of character traits. Second, I aim to draw out the conceptual connections between three familiar notions: *character traits, character flaws, and acting out of*
character. Finally, I aim to present a fresh interpretation of some of the experimental findings raised in recent discussions of character traits.

Before pursuing these aims, I want to briefly address a concern about whether my approach can be said to enjoy sufficiently broad philosophical appeal given the absence of any discussion of character-based ethical theories and the historical figures that feature prominently in them. Virtue theorists appear to be increasingly open to the prospect of developing a state-of-the-art moral psychology to champion their existing theories. My approach is complimentary to this general trend, and should, therefore, be relevant to proponents of character-based ethical theories. Furthermore, as Brandt (1970: 23) and others have pointed out, the nature of character traits and their tie to action is worth coming to grips with independent of the role they play in character-based ethical theories. Character trait ascriptions make frequent appearances in ordinary moral discourse, and the concept of a character trait features in various theories, including agency, moral blameworthiness and praiseworthiness, as well as “real” (or “whole”) self approaches to moral responsibility. Lastly, by freeing myself from certain theoretical and historical commitments, it is possible to offer a unique perspective that advances the debate in a different direction from those pursued at length in copious works produced over the last ten or so years (cf. Adams 2006: 119).

1. THE TRAIT-ACTION PROBLEM

In this section, I argue that each of the claims (1–3 above) are independently plausible. I contend further that denying any of one of them has unsatisfactory consequences, which is what

---

23 Daniel Russell (2009); Nancy Sherman (2010); and Christian Miller (forthcoming).
gives rise to the *trait-action problem* for proponents of dispositional theories of character traits. An adequate solution to the problem will, therefore, require an interpretation of one or more of these claims that would rationally permit a proponent to hold all three. I begin by fleshing out each thesis and noting its merits against those who would have us reject one or more of them.

1.1 THE DISPOSITIONAL THEORY OF CHARACTER TRAITS

As a first pass, and merely schematically, the dispositional theory of character traits is the view that a subject, S, possesses a character trait, T, if and only if S possesses a certain set of dispositions to think and feel in certain ways, as well as to say and do certain things in certain circumstances. In order to arrive at a less abstract characterization, I turn now to summarizing some of the sensible-sounding claims that have been made about the dispositional theory of character traits.

COUNTERFACTUAL CONDITIONALS

Perhaps the most distinctive mark of the dispositional theory of character traits is the link between traits and counterfactual conditionals via the dispositions associated with the trait in question. In illustrating this point, authors tend to draw our attention to various analogies between dispositions possessed by physical objects and those possessed by human beings. As Ryle says:

To possess a dispositional property is not to be in a particular state, or to undergo a particular change; it is to be bound or liable to be in a particular state, or to undergo a particular change, when a particular condition is realized. The same is
true about specifically human dispositions such as qualities of character (1949: 43).

For instance, David Lewis tells us that it is standardly thought that “a fragile thing is one that would break if struck [and] an irascible man is one who would become angry if provoked; and so on” (1997: 143). Brandt says, “Just as we might explain “x is soluble in water” as meaning “if x were placed in water it would dissolve,” so it is natural to propose...that “x is vain” amounts to a set of subjunctive conditionals, one of which might be “if it occurred to x that doing A would likely secure the admiration and envy of others, he would be strongly tempted to do A”’” (1970: 24–25). Accordingly, for every disposition (e.g., solubility), there is a corresponding activation condition (e.g., immersion in water) as well as a corresponding manifestation (e.g., dissolving) such that were the object in circumstances in which the activating state of affairs obtained, the manifesting state of affairs would also obtain. Importantly, neither state of affairs (activating or manifesting) need actually obtain in order for an object to possess the relevant disposition. This fact will continue to be relevant in this, and later sections.

It may seem as though I am being overly fastidious in fleshing out the dispositional theory, but the details do matter. Consider just one (potentially devastating) infelicity in the statement of the theory found in John Doris’s seminal work. He rightly notes that on the dispositional theory, trait-attributions are associated with a conditional, but he then wrongly goes on to gloss this as the material conditional: “If a person possesses a trait, that person will exhibit trait-relevant behavior in trait-relevant eliciting conditions” (2002: 16). There is a straightforward reason why the material conditional is not the sort of conditional used in the dispositional account; it falls under the heading of the problem of false antecedents. If my guinea
pig is brave, she will reach the summit, if on Mount Everest. This conditional is true because all of the antecedents are false. To give another example, a rubber ball that never encounters a certain activating state of affairs is fragile: if it is struck, it will shatter. It’s never struck (let’s suppose), so it is true on this reading that the rubber ball is fragile.

The principle of charity would have us interpret Doris’s statement more favorably. But the upshot is this: the proponent of the dispositional theory of character traits has every right to demand precision, for it is her theory that will be needlessly subject to counterexamples, among other things, that would doom her theory from the outset, were the proper statement of it not in place.

Of course, famously the simple counterfactual conditional analysis of dispositions faces its own set of problems, but this need not trouble proponents of the dispositional theory of character traits, for they are in a position to wait for the proper analysis to come along. Alternatively, these proponents have the option to relax the logical relation involved in their proposal. Instead of necessary and sufficient conditions, for instance, she might suggest that a person possesses a trait in virtue of certain counterfactuals concerning that person. The grounding relation, as it called, has been the saving grace for other failed analyses, and, yet it satisfyingly preserves what was initially plausible about those attempts (cf. Schaffer 2009).

HIGH-GRADE DISPOSITIONS

Even if the dispositional account can be made to work on logical and metaphysical grounds, it would still require further substantive refinements. In particular, Ryle raises a concern

---

about the analogy between dispositional properties possessed by objects and those possessed by humans. In this connection he says:

In discussing dispositions it is initially helpful to fasten on the simplest models...To be brittle is just to be bound or likely to fly into fragments in such and such conditions...These are simple, single-track dispositions, the actualizations of which are nearly uniform...Now the higher-grade dispositions of people...are, in general, not single-track dispositions, but dispositions the exercises of which are indefinitely heterogeneous (1949: 43).

There are two ways to interpret Ryle’s distinction between “single-track” and “high-grade” dispositions. On the first reading, for each trait there is single associated disposition, but indefinitely many manifestations of that disposition. On the second, for each trait, there are indefinitely many associated dispositions, the manifestations of which are collectively highly diverse or heterogenous.

I think the latter alternative makes more sense: a person possesses a particular trait in virtue of possessing sufficiently many of the dispositions associated with that trait. According to this “cluster” account of traits, there are dispositions to think and feel certain ways, as well as dispositions to do and say certain things, and these dispositions manifest in a variety of ways in certain circumstances.25 This account fits well with another of Ryle’s points:

When Jane Austen wished to show the specific kind of pride which characterized the heroine of ‘Pride and Prejudice,’ she had to represent her actions, words, thoughts, and feelings in a thousand different situations. There is no one standard

25 For similar proposals, see John Searle’s cluster theory of names (1958) and Eric Schwitzgebel’s dispositional theory of belief (2002).
type of action or reaction such that Jane Austen could say “My heroine’s kind of pride is just the tendency to do this, whenever a situation of that sort arose” (1949: 44).

Accordingly, any adequate account of character traits must include “high-grade” dispositions.

INTELLIGENT DISPOSITIONS

Human dispositions are “intelligent,” or so it often claimed. And, yet, it is not always clear what this claim amounts to. Part of the difficulty is, as Ryle notes, that there is a longstanding tendency to seek out some faculty, or set of “internal causes,” that exists over and above our dispositions and imbues them with intelligence (1949: 33). This propensity is especially pervasive in the Aristotelian tradition. Consider what Rachana Kamtekar says in this connection: “A virtuous disposition is a disposition to act and feel in particular ways in response to rational considerations; it is expressed in our decisions, which are determined through rational deliberation” (2004: 479). This is a complex statement, involving a commitment to several different relations. Rather than attempting to prize them apart, it will suffice for my purposes to argue that neither rational deliberation nor means-end practical rationality appear to be necessary or sufficient for a disposition to count as intelligent.

Consider, first, the necessity claim. Imagine a woman walking on a busy city-street. She is about to approach the corner of a busy intersection, when she notices that the person walking next to her is about to (apparently absentmindedly) walk into speeding traffic. She swiftly grabs his collar, and yanks him back from the onrushing cars. All of the following seem to be true of our heroine: she did not plan to save his life, she did not have a pre-existing desire to do so, and
she certainly had no time to self-consciously deliberate about whether to save his life or to reflect on the best strategy for doing so. And, yet, we admire the woman for her action, and think her relevant dispositions intelligent. Arguably, her awareness of her surroundings; her quick thinking; and speedy reaction time are all manifestations of intelligent dispositions, and these are, in turn, associated with certain of her traits, such as her conscientiousness, benevolence (or good-naturedness), and quick-wittedness. Importantly, moreover, the activation of these dispositions, as well as the dispositions themselves, do not require guidance from some additional rational faculty or act of deliberation.

I also claimed that rational deliberation and practical rationality are not sufficient for a disposition’s being intelligent. This is clear in cases in which emotionally intelligent dispositions manifest admirable actions. The soldier storming the town is sensitive to the danger at every turn. He is disposed to feel fear of a certain intensity in response to (possible) danger, but when this fear manifests, it is not so intense that it clouds his thought or dampens his reflexes. Likewise when the comedian performs successfully he is emotionally sensitive to his audience. Ryle describes a clown pretending to be clumsy: “He trips and tumbles just as clumsy people do, except that he trips and tumbles on purpose after much rehearsal and at the golden moment and where the children can see him and so as not to hurt himself” (1949: 33). I contend that these actions are, in part, the manifestations of emotionally intelligent dispositions. Thus, rational deliberation and practical rationality are not sufficient for a disposition to count as intelligent, a certain kind of emotionality is also a quality of intelligent dispositions.
WELL-MOTIVATED DISPOSITIONS

Another claim often found in the literature is that the dispositions associated with particular character traits must be “well-motivated.” This is not the idea that dispositions require an extra “push,” as it were, from a corresponding desire in order for the manifesting state of affairs to obtain, for dispositions are similar to motives in that “they incline without necessitating” action, to use Leibniz’s phrase (1956: 57). In this way, they do not need an extra push; they come with their own. Rather, the idea is that if a particular disposition is cultivated as a result of the person’s wayward or questionable motives, then such a disposition should not be associated with certain (positive) character traits. Brandt provides a familiar example to illustrate: someone who possesses a disposition to give to charity simply because he has a very strong desire “to be in the public eye, or improve his reputation, or to be elected to public office” is not generous (1970: 30). If these wayward motives are the foundations for certain of our dispositions, then those dispositions are not the ones associated with traits of character.

STABLE AND ENDURING DISPOSITIONS

The dispositions associated with character traits should also be stable and enduring. They need not be permanent, however. It is possible for a person to lose dispositions, and also to cultivate new ones. This might happen after a life-altering event or through training and new experiences. What the claim should not be mistaken for is the entirely different thesis that someone who possesses a stable and enduring disposition will also exhibit stable and consistent overt behavior. This is a topic I return to in the next section.
1.2 THE MASKING THESIS

Dispositions are our friends—at least most of the time. They are crucial to the process of coordinating sophisticated forms of behavior efficiently and automatically, often initiating fluent and appropriate, even admirable, actions. Unfortunately, in order to produce such results, dispositions need not be particularly nuanced or context-sensitive, and this can get us into trouble. I was once walking down my apartment hallway, when out of the corner of my eye, in a passing corridor, I saw my neighbor kneeling down with a roll of paper towel in his hand. Before I could fully process what was happening—he was in fact attempting to clean up a large amount of his dog’s diarrhea—I gleeful said “Hey! How’s it going?” This embarrassing misfire is just one example of how our dispositions can lead us astray: they sometimes manifest in inappropriate circumstances. The flip-side, of course, are those occasions when it is appropriate, even required, that our dispositions manifest, and, yet, they let us down by failing to do so. In these cases, the manifestations of our dispositions are “masked” or interfered with (cf. Johnston 1992). This section is dedicated to exploring this phenomenon.

When *masks* are present, as John Hawthorne and David Manley put it, they “interfere with the manifestation of the disposition without removing the disposition” (2005: 180). Michael Fara illustrates this phenomenon and just how prevalent it is when he says:

I’m disposed to go to sleep when I’m tired; but this disposition is sometimes masked by too much street noise. Cylinders of rubber are disposed to roll when placed on an inclined plane; but this disposition can be masked by applying a car’s brakes. A piece of wood in a vacuum chamber is no less disposed to burn
when heated than is its aerated counterpart [...]; but wood won’t burn if heated in a vacuum (2005: 50).

Accordingly, there appears to be no relation of dependence between something’s possessing a disposition, and the obtaining of its corresponding manifesting state of affairs. In fact, the authors, C. S. Jenkins and Daniel Nolan (2011), take matters a step further when they argue that it may even be impossible for a disposition to manifest, and, yet, this would not entail that the thing did not possess the disposition in question.

The existence and pervasiveness of masking is just as applicable to persons and their dispositions. I want to argue that even for traits, such as truthfulness, in which we typically think it important that its associated dispositions manifest, masks can interfere with the manifestation, and yet it does not undermine attribution of the trait to the person in question. Consider, first, what Robert Adams has to say in this connection:

Non-violence, like truthfulness, and more broadly, conscientiousness, looks like a virtue of perfect obligation. With the exception of a debated minority of cases in which violence, arguably, is justified, we are morally required to refrain from violence toward each other all of the time. And likewise with regard to truthfulness (2006: 124).

It is possible to interpret this passage as one about dispositions and their corresponding manifestations. On this reading, Adams appears to be suggesting that in order to be truthful, for example, one’s corresponding dispositions, such as the disposition to tell the truth, must manifest in nearly every circumstance in which the activating state of affairs obtain.
The passage is illustrative of two important points. First, Adams makes room for the phenomenon of masking when he claims that there are at least some exceptions: in order to be considered a non-violent person one’s corresponding behavioral dispositions may fail to manifest, on certain rare occasions. For instance, if one’s disposition to refrain from violence toward others were masked, as it likely would be in cases of self-defense or if someone were threatening one’s family.

The second point is a critical one: the passage above reveals that Adams is implicitly relying on a *ceteris paribus* clause. In other words, he assumes that the hypothetical set of persons in question are lucky enough to live in a society much like ours where certain freedoms and their corresponding obligations are generally respected. This, however, is a contingent matter: things could be very different, and we do not have to go to nearby possible worlds to see this. In China, people living in villages with corrupt leaders have been regularly sent to “work camps” for reporting abuses or corruption. If I were in this situation, and I did not know who I could trust, including government officials, neighbors, and possibly even family members, my disposition to tell the truth would fail to manifest on many occasions, and justifiably so. Similarly, if I was living in a society where violence was prevalent, in which my life, or the lives of my family members, were often in danger, my disposition to non-violence would frequently fail to manifest.

The upshot is that if we are more open-minded about our choice of examples, we see that a person’s disposition may *frequently* fail to manifest, and yet this alone does not entail that the person does not possess the corresponding disposition. Nelson Goodman compellingly articulates this familiar idea when he says:
Besides the observable properties it exhibits and the actual processes it undergoes, a thing is full of threats and promises. The dispositions or capacities of a thing—its flexibility, its inflammability, its solubility—are no less important than its overt behavior [...] (1954: 40).

Similarly, for truthfulness, non-violence, conscientiousness, and any other number of character traits a person might have. This point will loom large in a later section.

Of course, masking cases are not the only type of case in which a person might possess a disposition, while the manifesting state of affairs fails to obtain. It is also possible that the activating conditions for that person’s disposition never obtain. For instance, a person might be courageous even though they are never challenged, but were they challenged, they would rise to the occasion. Imagine a soldier who, in his first time in battle, is shot down by a sniper just as his platoon is preparing to charge into battle. The simple fact that his disposition failed to manifest does not entail that he is not courageous because it was never given a chance to manifest.

In this section, I argued that masking is a very real and pervasive phenomenon. But this is not to say that some philosophers have tried to offer dispositional analyses that make no room for it. Some views, such as Elizabeth Prior’s, take context-sensitivity into consideration. The context-sensitivity at issue, however, is typically interpreted as applying only to disposition ascriptions. What someone means when they say “The glass is fragile” is that “The glass is disposed to shatter into fragments in a highly specific circumstance C.” For this reason, David Manley and Ryan Wasserman (2008) call it “the getting specific strategy.” I set this approach to one side because here we will be concerned not so much with disposition ascriptions, but with
dispositional properties themselves.

1.3 THE EMPIRICAL THESIS

The dispositional theory of character traits combined with the masking thesis would, it seems, have us believe that although people possess character traits, they nonetheless rarely act as if they do—to put it somewhat crudely. But this conflicts with the natural thought that we seemingly justifiably rely on people to act in accordance with their character traits much of the time, and we are able to explain their behavior in terms of their character traits, as well as predict future behaviors, all with a relatively high degree of accuracy. Adams puts as fine a point as any on the central issue here when he reminds us how reliable we must be: If we were not able with any accuracy to predict people’s behavior on the basis of their traits, then, he asks, “how rational would it be for me to step into the crosswalk as a pedestrian?” (2006: 123). Accordingly, there appears to be a strong empirical connection between a person’s trait-relevant dispositions to act in certain ways in certain circumstances and the person’s performing the token actions which are the manifestations of those dispositions. The question is: what is the nature of this strong empirical connection? In the remainder of this section I attempt to articulate two possible answers to this question, and, in a later section, I attempt to reconcile one of these answers with both the dispositional theory of character traits and the masking thesis.

On one familiar view, the strong empirical connection is stated probabilistically. If a person possesses a disposition to perform certain actions in certain circumstances, that person’s disposition will manifest relatively frequently (cf. Brandt (1970: 31) and Doris 2002: 19). This view allows the possibility that the manifestation of a disposition can be masked, but it suggests
that masking is, in actuality, statistically rare. If true, it would also explain why we are reliable in our predictions of people’s overt behavior on the basis of their character traits: we have a statistically good chance of being right when we predict that someone will behave in accordance with their character traits. The problem with this view, however, is that the empirical data does not appear to bare out such a connection.

A second view—one which I defend in much greater detail in a later section—asks us to reconsider what role character traits should play in the explanation and prediction of action (and behavior). It suggests that we are satisfied with our predictions involving trait-relevant behavior even when they take the following form: either she will perform $X$; or she will not perform $X$, but she will have a good reason for failing to do so; or she will not perform $X$ and she will not have a good reason for failing to perform $X$, but her action can nonetheless be explained in some other way that is compatible with her possessing the trait in question. This view suggests that we can predict and explain the manifestations of traits and their associated dispositions not just in the direct way, such as truth-telling for truthfulness, but also by appealing to a more nuanced set of actions (behaviors) that tend to occur when a person’s disposition is masked. To illustrate the idea, consider Fara’s example above: he says he is disposed to go to sleep when he is tired, but the manifestation of that disposition is sometimes masked by too much street noise (2005: 50). If he is genuinely disposed to go sleep when he is tired, and is not instead a night-owl looking for an excuse to stay up, then we can expect him to do the following sorts of things: yawning, tossing and turning in frustration, eventually purchasing some earplugs or soundproofing for the windows, and so forth. The night-owl, by contrast, would likely not exhibit any of these behaviors. On this view, therefore, the strong empirical connection is not necessarily between
trait-relevant dispositions to act (behave) in certain ways in certain circumstances and their direct correlates in action. Rather, the connection exists between the trait-relevant disposition in question and a whole range of behaviors, which the person might exhibit even if the manifestation of their “direct” disposition is masked. I discuss some of the advantages of this view over the previous one in the third section.

Some philosophers dispense with the empirical thesis altogether. Advocates of this strategy are often working within the Aristotelian tradition. They contend that virtue and vice, and, therefore, positive and negative character traits are an ideal to be aimed at, not one typically realized by actual human beings. Adams, by contrast, suggests that “it is important to find moral excellence in imperfect human lives” (2006: 119). I could not agree more, which is why I will set this proposal to one side. Furthermore, a genuine solution to the trait-action problem will make it possible to maintain a commitment to all three of the following claims: the dispositional theory of character traits, the masking thesis, \textit{and} the empirical thesis.

2. PROBLEMS WITH EXISTING ATTEMPTS TO SOLVE THE TRAIT-ACTION PROBLEM

One way to try solve the trait-action problem is by adding a rider to the statement of the dispositional theory of character traits that specifies a certain probability or frequency with which the associated disposition must actually manifest in order for the person to possess the character trait in question. Masking is possible, on this view, and it actually occurs, but is statistically infrequent, so the masking thesis must also be amended. Furthermore, the empirical thesis is interpreted probabilistically: it states that there is a strong empirical connection between
possessing a certain character trait and the reliable manifestation of the dispositions associated with that trait.

The proposal has merits, the most important of which is that it allows someone to consistently hold the dispositional theory of character traits, the masking thesis, and the empirical thesis. In fact, the empirical thesis is entailed by the dispositional theory of character traits (once the aforementioned rider is added). It should be noted, however, that this last point is, to some extent, a double-edged sword for proponents of the proposal because a consequence of the entailment between the dispositional theory and the empirical thesis is that if the latter thesis is false (and the former is true), then most people fail to possess character traits. It is this particular vulnerability that has been the central topic of the literature on character traits for the last ten years or so.

By way of illustration, consider a segment from John Doris’s précis of his 2002 monograph, which contains a description of its central argument:

1. If behavior is typically ordered by robust traits, systematic observation will reveal pervasive behavioral consistency.

2. Systematic observation does not reveal pervasive behavioral consistency.


A “robust trait” for Doris just is the dispositional theory of character traits plus the frequency rider, so the above modus tollens argument challenges this particular version of the dispositional theory of character traits. Furthermore, a significant portion of the literature on character traits in recent years can be construed as a response to this argument.

26 The following is the definition of a robust trait Doris settles on: “If a person possesses a trait, that person will engage in trait-relevant behaviors in trait-relevant eliciting conditions with markedly above chance probability $p$” (2002: 19; emphasis in original).
I find the preoccupation with Doris’s argument puzzling. It is indeed a flaw of the “add a rider” solution to the trait-action problem that it is subject to the kind of argument Doris articulates. But what I do not understand is why this particular version of the dispositional theory of character traits should be our starting point at all, especially given that Richard Brandt had already—some thirty years earlier—offered independent reasons (decisive ones in my view) for abandoning the dispositional theory of character traits with the frequency rider. The proposed solution, in other words, looks like a non-starter for a proponent of the dispositional theory of character traits, so to frame the debate in a way that makes it seem as though the existence (or nature) of character traits hinges on it, dangerously verges on a straw-man argument.

That said, working within the constraints of the proposal, certain philosophers, such as Robert Adams, have gone some way toward making the view seem plausible despite its limitations. In this section, I discuss Adams’s arguments, as well as Brandt’s concerns about the theory of traits it presupposes, supplementing it with a few of my own worries about how this would fit into a proposed solution to the trait-action problem.

The view under discussion is what I will call “the frequency theory of character traits.” It says:

27 In this connection, both Brandt and Adams focus on what the former author calls “the direct disposition view.” Brandt defines it as follows: “for various trait-names a form of behavior typical of that trait can be identified (as talking for talkativeness), and that what it is for a person to have a certain trait is primarily for him to be disposed to behave in the correlated way, in certain conditions, relatively frequently. (Another way of putting it is to say that to have a certain trait is for the probability to be relatively high that the person will behave in the correlated way in certain conditions)” (1970: 31; emphasis in original). The reason I do not discuss the view under this heading and definition is because I think it fails to distinguish two different issues. The first issue is whether traits should be directly associated with certain dispositions to behave where the disposition in question has a kind of primacy in trait-attribution. The second issue, which is the focus here, is about building a condition into the dispositional theory of character traits that specifies the degree to which the associated dispositions must actually manifest in order for the person to possess the trait in question.
A person possesses a character trait just in case that person possesses a certain set of dispositions to think, feel, and act in certain ways in certain circumstances, *and these dispositions actually manifest relatively frequently in those circumstances.*

Adams has his own reservations with the view (2006: 120), but he is willing to go along with it at least with respect to some traits, such as truthfulness, non-violence, and conscientiousness. Here he suggests that in order for a person to possess one of these traits, that person’s dispositions must manifest, as he puts it in an earlier passage, nearly “all of the time” (2006: 124; emphasis in original). Accordingly, all one needs in order to undermine this claim is a single example of a person who possesses one of these traits even though it rarely manifests.

I offered two such examples in a previous section of the paper, but here is another. Imagine a young Jewish girl living in Nazi Germany who, through a stroke of luck, is mistaken for a gentile. 28 Naturally, the manifesting state of affairs for the dispositions associated with her trait of truthfulness would almost never obtain after this point even though the activating state of affairs would frequently obtain. In other words, when asked, she would not be truthful about her name, her origins or background, where she was from, her family members, her dietary practices, and so forth. Suppose, moreover, that after ten years of keeping up the ruse, she is eventually found out, and tragically murdered by the Nazis. We can suppose, further, that in her short life the dispositions associated with her truthfulness failed to manifest less than half the time. I contend that such a person can still be truthful despite the fact that the dispositions associated with her trait of truthfulness failed to frequently manifest. If this is right, then the frequency theory of character traits is inadequate even as an account of traits such as truthfulness. Its fatal

---

28 The example is intended to resemble the type of scenario portrayed in the film *Europa.*
flaw, moreover, can be located precisely in the addition of the frequency rider as a necessary condition on character trait possession.

The problems with the frequency theory do not end at counterexamples. This becomes clear in Adams’s discussion of generosity, when he says:

If we ask, “Is she a generous person?” we do not mean, “Does she show notable generosity on every occasion?” Nobody has the personal resources for that. It will be worth asking her help, and appropriate to praise her character in this respect, if she is disposed to act [and does in fact act] with notable generosity as much as 10 percent of the time” (2006: 123).

I want to take issue with what Adams says in the second sentence because, although I agree with his first statement, it seems equally true that if we ask, “Is she a generous person?” we do not mean, “Does she show notable generosity at least 10 percent of the time?” The passage, therefore, unwittingly illustrates Brandt’s observation that “we do not think of traits as statistical facts” (1970: 34). That is, when we attribute a character trait to someone, we do not first take-stock or quantify that person’s trait-relevant behavior.29

Brandt has several examples illustrating the point that we sometimes rightly attribute traits to people on the basis of a single piece of behavior. He imagines a boy who stands up to some stronger playground bullies (1970: 26). Once we have ruled out certain alternative explanations (e.g., that he did not realize the danger) we justifiably deem him courageous. I would add to what Brandt says that the reason we can reliably judge traits on the basis of a single behavior is partly due to the fact that paradigmatic forms of overt behavior are not the only

29 Another problem in the neighborhood is why 10% and not 11% or 9%? Any given percentage we provide appears to be arbitrary.
factors our trait attributions are based on. For instance, if on a first encounter with someone, I think they are being gratuitously hostile, although not in an immediately obvious way, I will point to a number of factors that justify my attribution, including his overt behavior; the things he actually says, but also what his utterances implicate; his “body language”; tone of voice, and so forth. Thus, our evidence base is often much more substantial than is often assumed in discussions of character trait attributions.

The upshot of this discussion is that the frequency theory of character traits does a poor job of explaining our practices of trait attribution. Our actual trait attributions do not rely on statistical patterns of behavior, but are often hypotheses (subject to revision), which are sometimes based in part on a single overt behavior, but supported by copious evidence regarding the manifestation of the person’s other trait-relevant dispositions.

And, yet, the tendency to associate character traits with certain overt behaviors is pervasive in the literature. I think this tendency could be a by-product of our language. Notice that some terms, such as “honest” and “faithful” are ambiguous: they have both a dispositional and an episodic meaning. For instance, the question of whether someone was honest on a particular occasion has a “yes” or “no” answer. If he lied on that occasion, then he was not honest on that occasion. This, however, should not be confused with the dispositional notion of honesty—a property a person could possess even if he failed to tell the truth on many occasions.

It is also true that philosophers are not always careful to distinguish between dispositions and overt behavior. Adams makes this category mistake in the above passage when he says that someone can have a disposition to act in a certain way 10 percent of the time. It is the action, which is the manifestation of the disposition that is appropriately said to occur a certain number
of times, not the disposition itself. To give another example, in a review of Adams, Doris does something similar. He first quotes Adams as saying that dispositions to behave in certain ways in certain circumstances are necessary for possessing certain character traits, but then he goes on to say: “[...] This seems obviously true, and I won’t discuss it further, pending a compelling—very, very compelling—argument to the effect that overt behavior is not necessary for virtue (2010: 141). Adams says dispositions to behave are necessary for possessing character traits (virtues), but Doris glosses this as the claim that overt behavior is necessary for possessing character traits. Of course, if one accepts the frequency theory of character traits, then the frequency of overt behavior is necessary for possessing character traits, and, yet, relevant to the substance of Doris’s claim, we have seen that there are, in fact, compelling reasons for rejecting this theory.

The proponent of the dispositional theory of character traits knows full well that solving the trait-action problem requires articulating the empirical thesis in such a way that she can consistently accept it, while also holding the masking thesis. The question is how to interpret the empirical thesis and, in particular, what the strong connection between traits and behavior consists in that does not rely on the notion of frequency? In the next section, I defend answers to these questions.

3. A NEW SOLUTION TO THE TRAIT ACTION PROBLEM

In the last section, I defended the following claim: what matters for a dispositional theory of character traits is not that a particular disposition has failed to manifest on a given number of occasions. Our concern should rather be—or so I argue in this section—what it is that interferes with the manifestation of a person’s trait-relevant disposition. This shifts the emphasis away
from the frequency theory and focuses instead on the source of the interference. I argue that the strong empirical connection between traits and action can be maintained on the grounds that the types of interference that come between traits and their associated actions are relatively circumscribed and that they are compatible with the manifestation of certain kinds of actions and behaviors even if not the paradigmatic ones. It turns out, then, that character traits do play an important explanatory and predictive role with respect to action and behavior, albeit a different and more nuanced one than is typically presumed.

I begin by discussing three ways in which the manifestation of a person’s trait-relevant disposition to perform a certain action in certain circumstances can be masked without undermining the attribution of the trait to that person. These are:

1. *Character adjustments:* when the manifestation of a person’s disposition to perform a certain action in certain circumstances is masked, but her failure to act is nonetheless justified.

2. *Acting out of character:* when the manifestation of a person’s disposition to perform a certain action in certain circumstances is masked, but her failure to act can be explained by the extenuating circumstances she faced.

3. *Character flaws:* when the manifestation of a person’s disposition to perform a certain action in certain circumstances is masked owing to certain isolated, but problematic beliefs or desires.

After explaining these three notions in greater detail, I will argue that they help the proponent of the dispositional theory of character traits to solve the trait-action problem.
3.1 CHARACTER ADJUSTMENTS

It is uncontroversial that if the manifestation of a person’s disposition is masked for morally justifiable reasons, then the fact that it was masked removes any ground for withdrawing the attribution of the character trait in question. What is not so uncontroversial is whether a certain set of experimental data illustrates this particular phenomenon. I have in mind here those experiments that sometimes fall under the heading of “mood-effect” experiments. I focus here on one of the more widely cited experiments.

In the 1970’s, the psychologists, Matthews and Canon, wanted to know whether high noise levels would influence their subjects’ helping behavior. Staged in a residential neighborhood, this experiment used passersby as participants. While walking along the sidewalk, subjects encountered a man (actually the confederate) holding some apparently heavy boxes. Several items were also lying on the ground, so that it appeared to the participants that he was not in a position to easily pick them up on his own. In one condition, participants were subjected to the loud noise of a nearby power mower running without its muffler. 10% of these participants helped the confederate pick up the items in this condition, whereas 20% helped in the control condition with ambient noise alone (1975: 574). John Doris and Stephen Stich (2005) claim that the subjects were “readily induced” to behave in morally problematic ways by “insubstantial” situational influences, such as the loud noise of a power mower (2005: 118–119: cf. Merritt, Doris, and Harman 2010: 356–357). I want to argue to the contrary that an examination of the details of the experiment reveals that although the manifestation of the subjects disposition to help was masked, it was for morally justifiable reasons.

---

30 To be more precise, Doris and Stich (2005: 118) offer the statistics for a slightly different variation of this experiment in which the confederate was wearing a shoulder-length cast on his arm. I will discuss the results of this experiment in a footnote below.
In discussing a related case, Adams notes (plausibly, I think) that we should distinguish between perfect and imperfect obligation. “A perfect obligation,” according to Adams, “is one that is violated if one fails in any single case to behave in a particular way” (2006: 145). “An imperfect obligation,” he continues, “is satisfied if one “does enough” of the relevant sort of thing” (2006: 145). Adams thinks this is what explains the results in cases similar to the Matthews and Canon experiment. Discussing a related experiment, he says:

Both helpers and non-helpers probably regarded helping in such a situation as a non-obligatory kindness rather than a matter of strict and perfect obligation (2006: 145).

While I think Adams is correct on this point, this cannot be the whole story. For, as his phrase “non-obligatory kindness” suggests, kind people tend to help in these contexts even if they consider the action to be non-obligatory or supererogatory. That is, to help in situations like these is one of the marks of being a kind person. Accordingly, we want to know why so few people helped in both the ambient control condition (only 20% helped) and in the noisy condition (only 10% helped).

As just noted, kind people are not usually dissuaded from helping on the ground that the action is non-obligatory, but they are rightly dissuaded from helping someone when the action is non-obligatory and when the person who seems to need help, nevertheless, appears to not want help. Were someone to help in such a circumstance, we would be more inclined to describe her action as intrusive than as kind. Accordingly, it is reasonable to wonder whether the participants in the Matthews and Canon experiment had some reason to believe that the confederate did not want their help. In their description of the experimental design, Matthews and Canon hint at an
answer to this question, for they say: “at no time did [the confederate] glance toward the subject or in any way indicate that he wished assistance” (1975: 574). A plausible hypothesis, then, is that in controlling for the confederate’s behavior in the way they did, Matthews and Canon secured a non-helping result in both conditions, simply because, when someone does not indicate that one wants assistance, this is easily interpreted as indicating that one does not want it.\footnote{Other researchers have not been so careful to control for the confederate’s behavior. In a similar experiment, for instance, Isen and Levin (1972) do not report having controlled for the confederate’s behavior. Yet, as Miller has noted, the results of this experiment have not always been successfully replicated (2009: 148–149). The failure to control for the confederate’s behavior is, I think, one plausible explanation of this fact.}

My evidence for this claim comes in part from reflecting on my own phenomenology in interpersonal encounters of this kind (and I would be surprised if I am in idiosyncratic in this respect). It strikes me that the confederate’s “body language” is precisely the type of behavior that sometimes inhibits me from helping others. And rightly so, for we partly gauge whether it is appropriate to intervene in the affairs of others especially in non-obligatory contexts like these by responding to the subject’s verbal behavior and “body language.” Eye contact, asking for help, and smiling usually welcomes help, while no eye contact, failing to ask for help, scowling, or having no particular facial expression at all often indicate the desire not to be helped (or, minimally, they send “mixed signals”). Such factors, moreover, are not trifling—or “insubstantial,” as Doris and Stich would put it—insofar as they guide successful interpersonal encounters by encouraging appropriate responses to otherwise ambiguous and epistemically non-ideal circumstances.

If I am correct in thinking that the majority of participants in the Matthews and Canon experiment were dissuaded from helping in part because they judged helping to be non-obligatory and because of the confederate’s unwelcoming body language, then we should not at
all be surprised that the majority of people failed to help in both the ambient control condition and the noisy condition. For, on such an interpretation, what masked the manifestation of their disposition to lend a helping hand was not some trivial factor, but the justifiable consideration that the confederate did not seem to want help. Nevertheless, I have not yet accounted for the discrepancy between the 20% who helped in the control condition compared with the 10% who helped in the noisy condition. But there is a fairly straightforward explanation for this: subjects may have thought that the noise would make it difficult to ask the confederate if he needed help, or to try to communicate their intention to help—something one might feel obliged to do if one suspects that the person does not want help.32

3.2 ACTING OUT OF CHARACTER

I argue here that when a person faces extenuating circumstances, and, as a result, the manifestation of her trait-relevant disposition to perform a certain action in certain circumstances is masked, a person may act out of character (or contrary to her character traits), but this does not entail that she is character-less (or that she fails to possess the character trait in question). Stanley Milgram’s obedience experiments, I contend, serve as an illustration of this point.

32 In one variation of their original experiment, Matthews and Canon have the confederate wear a shoulder-length cast on his right arm. In this version, 80% of the participants helped in the control condition, and 15% helped in the noisy condition. The striking disparity in the data seems to reveal that a trivial factor (viz., the loud noise of the power mower) had a fairly severe impact on their morally relevant behavior. I will argue, to the contrary, that it was not a trivial factor that masked the manifestation of their trait-relevant behavioral disposition, so this version of the experiment does not undermine the attribution of kindness to them either. Let’s assume that the confederate’s shoulder-length cast changes the moral valence of this experiment so that participants now have a perfect obligation to help the confederate rather than an imperfect one. This, by the way, would explain why so many participants helped in the control condition. Participants in the control condition may have thought that even though the confederate’s body language was telling them that he did not want help, his seeming impairment (viz., his “broken” arm) was telling them that he needed help in a way that gave them a perfect obligation to help. Thus, the majority of people (80% to be precise) behaved just as we would expect them to: when no masks were present, kind people behaved kindly. Why, then, did so many participants fail to help in the noisy condition? For the very same justifiable reason that they failed to help in the other experiments: whether it is appropriate to help depends on the possibility of open communication with the person being helped—at least in non-emergency contexts such as these.
In the original version of the experiment, 65% of the subjects completed the experiment by issuing the highest voltage on the “shock generator,” 450 volts, to an unwilling recipient in response to an experimenter’s requests (Milgram 1974: 35). (In fact, the recipient was not receiving any shocks, and both he, along with the experimenter, were confederates.) A common reaction to the Milgram experiment is to condemn the obedient participants, and to confidently assert that, had one been similarly placed, one would have resisted the experimenter’s demands. From this perspective, the claim that Milgram’s participants faced extenuating circumstances has the air of an unwarranted “apology” of their morally unacceptable actions. To make matters worse, Milgram likened his obedient subjects to perpetrators of the Holocaust. As a consequence, were we to take Milgram’s comparison seriously, then someone who held that Milgram’s participants faced extenuating circumstances would also appear to be an apologist for the Nazis. Elsewhere I have addressed the comparison, as have others, between Milgram’s participants and perpetrators of the Holocaust, and the consensus appears to be that such an analogy is unfounded (cf. Leonard Berkowitz 1999). I will not rehearse those arguments here. Instead, it will suffice for my purposes to argue that the reason people standardly condemn Milgram’s participants as pathetically acquiescing to the experimenter’s demands is either because they have relied on caricatures of the experiment, or because they have failed to appreciate certain pertinent details of the experiment and its design, as is I think the case in many discussions of it in philosophy and psychology.

As early as the first published article on the results of the original “base line” experiment, Milgram hints at a reason for thinking that the circumstance his participants faced had to be at least somewhat exceptional. In this connection, Milgram reports the following:
In a large number of cases the degree of tension reached extremes that are rarely seen in sociopsychological laboratory studies. Subjects were observed to sweat, tremble, stutter, bite their lips, groan, and dig their fingernails into their flesh. [...] Full-blown, uncontrollable seizures were observed for 3 subjects. On one occasion we observed a seizure so violently convulsive that it was necessary to call a halt to the experiment (Milgram 1963: 375).

These are not ordinary behaviors. The natural inference to draw, therefore, is that the circumstances were also not ordinary. In particular, it is difficult to reconcile these extreme emotional and physiological reactions with what some authors have described as the “polite request” of the experimenter (Merritt, Doris, and Harman 2010: 357). I want to argue that, although his methods were subtle, the experimenter used bullying tactics, specifically, he played a role in triggering the intense emotional reactions described above, and then he took advantage of their compromised state. This is the central reason why the majority of Milgram’s participants complied with his demands despite being unwilling to do so.

The upshot of this discussion is that it is possible to act out of character in extenuating circumstances, and although the nature of these circumstances do not fully justify the actions performed in them, they should at least mitigate the extent to which we condemn the person for performing the actions in question. More to central point, we have reason to believe that the various reactions described above were manifestations of their kindness and non-violence.
3.3 CHARACTER FLAWS

Someone has a character flaw when she possess a wayward belief or desire that masks the manifestation of the person’s trait-relevant dispositions. A wayward belief or desire is one that is irrational or is *within certain limits* morally problematic. The qualification in italics is crucial, for if a belief or desire is morally problematic enough, it could be character undermining, not simply a flaw of one’s character. The best way to illustrate this idea is through examples.

Consider, first, an example of a wayward desire. Imagine a woman who has been married to the same man for ten years. She is not happy in her marriage, which is, in part, the by-product of an unexpected pregnancy. Over the years it has become clear to her that they are not compatible. Things are complicated, however. They have two children together, and one of them has a physical disability that requires extensive parental effort, and she gave up the chance of a career in order to raise her children. Although she increasingly finds her husband’s personality unbearable, he is nonetheless a caring provider for her and their children. One day, however, she becomes romantically involved with someone else and has an affair. She then meets another man and cheats again. Suppose now that she falls in love with the second man, marries him, and lives the rest of her days in a fulfilling marriage (forty-plus years). She is never unfaithful (in the episodic sense) to her second husband. Is this woman a faithful person (in the dispositional sense)? My inclination is to say that she is. It is true that she had an array of wayward desires, not just romantic in nature, but, also, the desire to be out of her marriage, and that these desires interfered with the manifestation of the dispositions associated with her faithfulness, and, yet, these desires and the fact that they mask her disposition to refrain from engaging in an extramarital affair do not, in my view, speak to the core of her character.
Articulating precisely what makes her wayward desire one that is not character undermining is no doubt challenging. Presumably, however, the circumstances that lead her to enter the marriage are relevant. Suppose further that her parents gave her an ultimatum: marry him or we will not help support you and the child. Another consideration relevant here concerns the norms and expectations of a healthy marriage. Her desire is born of an unhappy marriage, not her character. Moreover, in isolation the desire for a loving and romantically inclined partner is a healthy and reasonable one. Obviously, these considerations do not justify her desire to have an affair, or her acting on that desire. Nevertheless, given the context, we can understand her desire, and we should, I think, be reassured that her other values, beliefs, and desires are those in virtue of which a person is faithful in the dispositional sense, and her actions both in her tens years of the first marriage and in the forty years of the second are a testament to her fidelity.

Consider, now, a case of a wayward belief. In one of a series of experiments, said to demonstrate what is often called the “bystander effect,” psychologists Latané and Rodin (1969) discovered that when subjects were alone, 70% of them came to the assistance of of woman after hearing what sounded like a loud crash on the other side of a curtain, followed by the woman’s cries of pain (actually tape recording). By contrast, 7% came to the woman’s assistance when accompanied by an unresponsive fellow participant (actually a confederate). In these cases, a wayward belief appears to be partially responsible for the participants failure to respond appropriately. In particular, the bystander’s lack of action appears to influence the participants’ belief in one of two ways: either they come to believe that the woman does not need help or their credence in their belief that she needs help is significantly reduced to the point of being inefficacious. The question, again, is not one of justification. Clearly, neither the participants’
omission nor their belief (or credences) is justified when they fail to help in the presence of a bystander. The question is whether interference from of a belief of this kind is character undermining. In other words, is the fact that most of us are susceptible to the bystander effect a reason to think that most of us are not helpful (in the dispositional sense)? I say no.

First, the effect is highly circumscribed and isolated from the subjects network of beliefs. Contrast this belief, for instance, with a virulently racist belief that would be liable to infect other beliefs and desires, as well as interfere with the manifestation of a number of the subject’s trait-relevant dispositions (if in fact such a person possessed positive moral character traits at all). Second, as Cialdini (1993) points out, there are reliable techniques, such as asking for help, that easily overcome the effect. This ties into the idea that susceptibility to the bystander effect really is just a “flaw” that can fairly easily be eliminated. Third, questioning one’s judgment in the face of dissent is often epistemically commendable. In bystander effect cases, however, it backfires.

I make no claim to having provided a complete account of character flaws. Such a topic is deserving of its own treatment. Instead, my aim has been to motivate the idea that character flaws, which are not themselves character undermining, do exist, and that a complete account of character traits would need to be able to account for them.

I turn now to the issue of the trait-action problem, and how to resolve it. The proponent of the dispositional theory of character traits can maintain that masking is possible, and even that masks are in fact pervasive, while also holding that there is a strong empirical connection between character traits and action. This is because even when the manifestation of a person’s trait-relevant disposition is masked we now have a way of determining whether or not the source of the interference is character undermining. Furthermore, in such cases the person does typically
exhibit certain kinds of predictable behavior. For instance, most of Milgram’s subjects underwent severe distress, and we would be surprised if they had not. Milgram describes another subject, Bruno Batta, who, without any display of emotion, complied with the experimenter’s demands even when he had to force the recipient’s hand down onto a shock plate (1974: 45–46). Character traits and the phenomenon of acting out of character help to explain the palpable differences between Batta and the typical participant.

The final question I will consider is: in what sense, if any, have I offered a theory of character traits? After all, I have not offered a positive analysis. This is exactly right, and it is my view that proponents of dispositional theories of character traits have been lead astray in their attempts to offer such all-encompassing accounts. A theory of character traits is, in my view, much too complex to be captured in a single bi-conditional.

I propose a different alternative that can be illustrated by drawing a theoretical parallel to standard approaches to moral responsibility. These theories often take as their starting point the distinction between exemptions and excuses. On the exemptions side, certain adult human beings (and arguably children and non-human animals) are exempt from moral responsibility attributions because they fail to qualify as moral agents. The theorist must then provide a theory of moral agency or an account of when someone qualifies as being a member of the moral community. On the excuses side, we can distinguish between exculpating and mitigating excuses. An exculpating excuse gets one off the hook, as it were, because the performance of the action was not in fact morally unjustified. With respect to mitigating excuses, a person is still blameworthy for performing the action, but less so because, although the action was morally wrong, it was done in extenuating circumstances. The upshot is that questions of moral
responsibility are, roughly speaking, reduced to questions about exemptions and excuses, rendering the answers to those questions more tractable.

Similarly, I want to suggest that questions about character traits can be reduced to questions about exemptions and excuses. On the exemption side, certain adult human beings are exempt from character trait attributions because they lack certain capacities or fail to meet certain other conditions that would make it possible for them to cultivate the dispositions associated with certain character traits. The task would then be to specify what those capacities or conditions are. On the excuses side, character adjustments, acting out of character, and character flaws parallel the distinction between exculpating and mitigating excuses. Character adjustments are exculpating, while acting out of character and character flaws are mitigating, which is to say that someone continues to possess a character trait in the presence of these exculpating or mitigating factors. When looked at in this light, there is much work to be done for a theory of character traits. For instance, I did not even begin to address the issue of character trait exemptions in this paper. Nevertheless, the proponent of the dispositional theory of character traits can look forward to the potential for such progress.
REFERENCES


Latané, Bibb and Nida, Steve. "Ten Years of Research on Group Size and Helping."


