© Copyright by Simon Cullen, 2015. All rights reserved.
Abstract

According to the theory developed here, people’s moral attitudes play a fundamental role in determining whether they perceive the cause of an action to lie more within an agent or more in the situation in which the agent finds herself. I show how this view predicts surprising patterns in the attribution literature and I present new empirical studies in its support. I argue that the philosophical notion of self-disclosure (familiar from discussions of moral responsibility) and the person/situation distinction (familiar from social psychology) both pick up on the same bit of underlying folk psychology. This underlying folk psychology is plausibly understood as a manifestation of psychological essentialism—the pervasive cognitive tendency to locate hidden, causally active essences in a wide variety of entities. I consider some implications for theorizing about moral responsibility. Lastly, I apply the mismatch theory to better understand a fascinating feature of our thinking about how determinism and luck bear on responsibility, and I present new studies that confirm the mismatch theory’s surprising predictions in this area.
Acknowledgments

I owe sincere thanks to many friends and fellow graduate students, but Robin Dembroff, Yoaav Isaacs, Adam Lerner, Noel Swanson, and Vlad Ungureanu, deserve special mention. Their feedback has been invaluable. I’m especially thankful to Eva van der Brugge and Brennan McDavid: in addition to the hours upon hours of philosophical discussion I shared with them as Resident Graduate Students at Forbes College, they have also been absolutely magnificent friends to me. Grad school would not have been the great pleasure it was without them.

I have pretested far too many experiments on my students, advisees at Forbes, and many of the Princeton undergraduates who attended the Forbes Philosophy Table. I’m grateful to them for helping me not to stray too far from how people who aren’t trained as professional philosophers actually think about things.

Shamik Dasgupta helped me to think more carefully about many of the topics discussed here, and about how to present them most effectively. He has also convinced me that they have exciting applications I hadn’t even thought to consider. I can’t wait to work on these with him.

I am grateful to Johann Frick for getting me to think more deeply about modal conceptions of the person/situation distinction, and about how my views bear on questions concerning circumstantial moral luck. At many points I’ve found myself thinking back to questions Johann raised in a few extremely helpful discussions.

Michael Smith and Frank Jackson have supported me throughout my time at Princeton, and I owe them both a special debt of gratitude. Along with Gil Harman, they also provided very helpful comments on the materials that eventually came to form the basis of this dissertation.

English cannot possibly express how interplanetarily grateful I am for my advisers, Sarah-Jane Leslie and Gideon Rosen. I couldn’t have asked for a more
supportive, generous, or sagacious duo. They have been really nice to me. They have also carefully commented on so many versions of this dissertation, I’m sure they understand it better than I do. And while I haven’t been able to answer all of their criticisms, every page has benefited from their feedback.

I am grateful to many people outside of the Princeton philosophical community, but to none more than Joshua Knobe. Were it not for his warm encouragement and infectious enthusiasm, I would never have dedicated myself to the questions discussed here. Josh has given me direction from afar, and always with a light touch, but his influence has been immense.

I am also grateful to Angela Smith and Dana Nelkin, who both provided incisive comments on the material that forms the basis for the first two chapters.

I am thankful for the many helpful comments I have received from audiences at Princeton, Yale, NYU, and UCSD.

This dissertation is dedicated to my mother, my sister, my friends back home, and in memory of my father.
# Contents

Abstract ii

Acknowledgements iii

List of Figures viii

List of Tables x

Introduction 1

A conceptual question .............................................. 1

The mismatch theory, in brief .................................... 3

Scope and outline ................................................... 4

1 The modal theory 8

1.1 Motivating a modal approach ................................. 8

1.1.1 Three classics of situationism ............................. 12

The Good Samaritan .................................................. 12

Bystander apathy ..................................................... 13

The Milgram experiment ............................................. 14

1.1.2 Situational causation and self-disclosure ............... 17

1.2 Counterexamples to the modal theory ................. 21

1.2.1 Amending the modal theory ......................... 25
1.2.2 Individual differences ........................................ 28
1.3 Conclusion ...................................................... 31

2 The mismatch theory ............................................. 33
2.1 Extant evidence for the mismatch theory ..................... 36
  2.1.1 The Actor-Observer Asymmetry ............................. 36
  2.1.2 Intergroup attributions ...................................... 38
2.2 Modal explanations of attributional judgments ............... 41
2.3 Two new studies supporting the mismatch theory ............ 44
  2.3.1 Study 1: Mark’s move ....................................... 44
    Method .......................................................... 45
    Results .......................................................... 47
    Discussion ..................................................... 48
  2.3.2 Study 2: Kate’s pregnancy .................................. 50
    Method .......................................................... 50
    Results .......................................................... 52
    Discussion ..................................................... 54
  2.3.3 Ego-syntonic/Ego-dystonic ................................. 54
  2.3.4 A Watsonian challenge ..................................... 58
  2.3.5 The upshot for self-disclosure .............................. 60
2.4 Interpreting the mismatch theory .............................. 61
  2.4.1 The need for an account of moral valence ............... 62
  2.4.2 Psychological essentialism .................................. 63
    An essence-causal account .................................... 65
    An essence-congruity account ............................... 66
2.5 Conclusion ...................................................... 69

3 Attribution, Determinism, Luck ............................... 71
List of Figures

1.1 Comparison of the worlds considered by the simple and amended modal theories. .................................................. 26

2.1 The relationship between self-disclosure judgments and person/situation attributions according to the modal theory. .................. 33

2.2 The relationship between moral, modal, and self-disclosure/ attribution judgments according to the mismatch theory. ............... 35

2.1 Relationship between effect size (I-E) and inverse standard error separately for explanations of positive and negative events ............ 39

2.1 Study 1: Plot of attitudes to gay men vs. overall inclinations to situational attributions. .................................................. 49

2.2 Study 2: Plot of attitudes to abortion vs. overall inclinations to situational attributions. .................................................. 53

3.1 The interaction of luck and valence on responsibility ratings. ....... 86

3.2 The interaction of luck and valence on self-disclosure ratings. ...... 86

4.1 Study 1: ATLG scores vs. time taken to complete the survey (complete data set). .................................................. 110

4.2 Study 1: attitudes to gay men vs. situational attributions (complete data set). .................................................. 111
4.3 Study 1: How the same data looks when subjects who completed the
survey in under 2:15 are excluded. ........................................ 111
4.4 Study 2: attitudes to abortion vs. situational attributions (complete data
set). .................................................................................. 112
4.5 Study 2: How the same data looks when subjects who completed the
survey in under 2:00 are excluded. ............................ 112
List of Tables

1.1 The learner’s schedule of protests in Milgram’s obedience studies. . . . . 16
1.2 Prozi dialogue from Milgram’s “Obedience to Authority . . . . . . . . . 19

2.1 Study 1: Correlations of attributions with attitudes to gay men and 
political conservativeness. . . . . . . . . . . . . . . . . . . . . . . . . . . . 49
2.2 Study 2: Correlations of attributions with attitudes to abortion and 
political conservativeness. . . . . . . . . . . . . . . . . . . . . . . . . . . . 53

4.1 Descriptive statistics for survey completion times. . . . . . . . . . . . . 110
4.2 Study 1: Results of excluding only those subjects who failed at least one 
comprehension check and were in the top quartile for survey completion 
speed. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
4.3 Study 2: Results of excluding only those subjects who failed at least one 
comprehension check and were in the top quartile for survey completion 
speed. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
4.4 Study 2: Correlations for complete data set including subjects who failed 
speed and comprehension checks. . . . . . . . . . . . . . . . . . . . . . 114
4.5 Study 1: Result of excluding subjects who failed at least one comprehen-
sion check. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 114
Introduction

A conceptual question

John is normally stingy. Indeed, his avarice runs so deep he has left a family member to financial ruin when he could have easily afforded to help. Tonight, however, after watching his football team win the championship, John leaves a generous tip at his local sports bar. Does the cause of John’s generous action seem to lie within John himself, or is it more appropriate to say that his action was somehow caused by the situation in which he found himself?

This question points to a mystery in folk psychology and in social science. In a wide range of cases, people agree about which actions are caused by the person and which by the situation, but a moment’s reflection is enough to see that this distinction is anything but clear. John wouldn’t have given a generous tip had his team lost, so his action does depend counterfactually on an aspect of the situation, but this alone cannot explain the sense in which his action is situationally caused. For all action is the product of both internal dispositions and situations. The explanation “John gave a generous tip because his team won” implies that John has an internal disposition to tip generously when, for example, he is in a particularly good mood. Situations can only influence action in virtue of actors’ internal dispositions to respond to them. The epidermis is not, as many social psychologists have assumed, “a special boundary that separates one set of ‘causal
forces’ from another” (Gilbert and Malone, 1995, p. 21). But then what underlies our judgment that the cause does not lie within John himself?

This question is not only interesting in its own right; it is also important to social science. A rich tradition in social psychology attests to the intuitive distinction between actions which are attributable to a person and those which are attributable to the situation in which she acts.\(^1\) This research is thought to teach that, more often than not, even detailed knowledge of a person’s dispositions is of little help in correctly accounting for her behavior: the real driver of behavior is not the person, it is the situation—intuitively unimportant and often little noticed. There is, however, no widely accepted account of this distinction and social scientists’ writings often provide little help.

Consider, for example, Philip Zimbardo’s claim that the Stanford Prison Experiment was “assessing the extent to which the external features of an institutional setting could override the internal dispositions of the actors in that environment” (2011, p. 195). On a literal-minded reading, this actually doesn’t make much sense. For the external features of an institutional setting only get to affect behavior in virtue of actors’ internal dispositions to react to those features. But claims much like Zimbardo’s are made all over this area of social psychology, and I think we have to admit that they do appear, at least initially, to make sense. To take one more example, consider Ross and Nisbett’s claim that Stanley Milgram’s (1963) results do not provide “evidence that people are disposed to obey authority figures unquestioningly—even to the point of committing harmful and dangerous acts” (2011, p. 59). Ross and Nisbett cannot mean that Milgram did not reveal his subject’s dispositions; for if Milgram’s subjects were not, in the literal sense of the word,

\(^1\)In psychology the word *attribution* can refer to either (1) how people explain actions or (2) the causes people attribute to actions, or, as is most common, to both (1) and (2). This is not surprising given the close connection between causal and explanatory thought and talk; indeed, one well-known philosophical analysis of explanation holds that to explain something simply is to provide information about its causal history (Lewis, 1987). I haven’t found the ambiguity to be a problem, so I haven’t gone to any length to avoid it, but perhaps it is worth flagging.
disposed to obey to the point of committing harmful acts, Milgram would never have found his famous result. We should conclude that statements like these have another, less literal-minded, more intuitive, reading. A philosophical analysis of the person/situation distinction should make this reading available.

The mismatch theory, in brief

The view defended here is that people explain actions in situational terms because (and to the extent that) they suspect a mismatch in the moral valence of the action and the moral valence of the person. That is, people judge that an action is explained by the situation because they judge that a good person is doing a bad action, or a bad person is doing a good action. If I am right, one of the central distinctions of social science isn’t really scientific at all. If there is a person/situation distinction, it’s a normative distinction.

This account may also help us to understand the philosophical notion of self-disclosure. Many theorists have been attracted by the idea that we are responsible only for those actions which are self-disclosing—i.e., for those actions that reveal our true, deep, or real selves (e.g., Frankfurt, 1971; Watson, 1996; Smith, 2005). According to such deep self theories, judgments about self-disclosure provide part of the nonmoral basis for judgments about praise and blame. These philosophers differ on many important points, but they all agree that the relation of self-disclosure is supposed to capture something about an agent’s psychology; it’s supposed to be the kind of thing a naturalistic scientist might describe; the kind of thing that might serve as an input to a moral judgment.

At least so far as it is understood as a claim in moral psychology, I argue that this gets things exactly the wrong way round. On my view, the principles that we rely on to determine whether an action is self-disclosing are the very same principles that drive us toward person or situation attributions. Thus, according to
the mismatch theory, we judge that actions are self-disclosing because we judge that
the moral valence of the agent matches the moral valence of her action. If I am right,
people’s judgments of self-disclosure do not track an action-theoretic relation and
cannot provide a nonmoral basis for judgments about moral responsibility. They
are thoroughly moral to begin with.

Scope and outline of this dissertation

One might try to provide a metaphysical account of the difference between person-
and situation-caused actions—but that is not the project of this dissertation. I am
interested in the psychology of attribution; I want to understand the principles
that actually lead people to explain actions in more or less agential or situational
terms. I aim, first, to develop an account of these principles; and second, to use this
account to shed light on the person/situation distinction, the folk psychological
notion of self-disclosure, and some of the many connected questions that arise when
theorizing about responsibility.

Chapter 1 begins by clearing the way for the mismatch theory. Given the lack of
detailed philosophical work on this topic, this involves developing explicit alterna-
tive accounts of the person/situation distinction. The most promising alternative I
develop is the modal theory—an account inspired by Lewis’s (1986) counterfactual
analysis of causation. The modal theory holds that we use the person/situation
distinction to indicate how we believe an agent would behave across a range of
relevantly similar circumstances. Again, consider John. The modal theory says
that we explain his action in terms of situational factors because we judge that in
relevantly similar circumstances (e.g., in circumstances in which his team loses the
championship), he would have left his ordinary stingy tip. This view has many
theoretical virtues and elegantly models our judgments about many classic situa-
tionist experiments. However, I develop a class of counterexamples that strongly
suggest it cannot be correct. The chapter concludes by extending this critique to a closely related account inspired by strict conditional analyses of counterfactuals, and analysis of variance models of attribution that have been influential in social psychology (e.g., Kelly 1967, 1973; Pruitt and Insko, 1980).

Chapter 2 articulates and defends a radically different interpretation of the distinction as it is employed by social psychologists and the folk—*the mismatch theory*. On my view, the person/situation distinction does not concern the causes of actions or even the psychologies of actors. Rather, I argue that we provide situational attributions for actions because (and to the extent that) we suspect a mismatch in the moral valence of the action and the moral valence of the actor.

In the first part of Chapter 2, I argue that the mismatch theory offers a principled explanation for deep patterns in the intergroup attribution literature (e.g., Hewstone and Ward, 1985), and then show that it generates an extremely surprising prediction. Jones and Nisbett (1972) famously hypothesized that “actors tend to attribute the causes of their behavior to stimuli inherent in the situation, while observers tend to attribute behavior to stable dispositions of the actor.” This hypothesized “actor-observer asymmetry” quickly came to play a central role in social psychology. It is simple to show that the mismatch theory predicts that there is no actor-observer asymmetry. Rather, the classic asymmetry should typically appear only when actors are explaining morally bad actions. When they explain morally good actions, the asymmetry should reverse direction. Happily for the mismatch theory, this prediction has recently been confirmed by a surprising and authoritative meta-analysis (Malle, 2006).

The second part of Chapter 2 presents the results of empirical studies I designed to test the mismatch theory more directly. I show that moral attitudes towards homosexuality and abortion predict both attributions and self-disclosure judgments. I argue that deep self theories presuppose a notion of self-disclosure that does
not correspond to anything in folk psychology. If the mismatch theory provides the correct account of self-disclosure, then there simply is no action-theoretic relation to analyze in the vicinity of the folk concept. I argue that judgments about whether agents’ actions reveal their true selves cannot serve as a non-moral basis for judgments about responsibility and should play no role in normative theorizing.

The chapter concludes by sketching an account of the content of judgments about agents’ moral valences, and locates the mismatch theory in an established approach to social cognition. I show how the phenomena of attribution might be understood as a manifestation of psychological essentialism. Psychological essentialism, as distinct from philosophical essentialism, holds only that human beings have a pervasive cognitive tendency to posit hidden, causally active essences in humans and animals. There is now copious evidence that dispositions play two importantly distinct roles in folk psychology (e.g., Gelman, 2003; Gelman and Ware, 2012; Sobel et al., 2007; Leslie, 2013). First, folk psychology posits superficial dispositions which vary within kinds, are malleable, and unrelated to development. Second, it posits essential dispositions which are stable across time, and important to kind-membership and development. This folk distinction between essential and superficial dispositions plausibly underlies our judgments about both the person/situation distinction and self-disclosure. I argue for an interpretation of the mismatch theory according to which we are inclined to judge that an action is situation-caused and not self-disclosing to the extent that the perceived moral valence of the action differs from the perceived moral valence of the agent’s essence. On this view, the contents of self-disclosure judgments concern whether actions are appropriately related to essences, and our judgments about agents’ moral valences aim to describe the moral quality of their essences.

Finally, in Chapter 3 I turn to a puzzling asymmetry in our intuitions about when responsibility is compatible with determination: determinism seems to un-
dermine our responsibility for bad deeds far more readily than it undermines our responsibility for good deeds. Philosophers have provided several competing accounts of this asymmetry, but none can explain the full diversity of the phenomena. The mismatch theory suggests a radically different account of the interaction of responsibility and determinism: normative considerations influence our responsibility judgments because they deeply influence how we explain actions. Normative considerations can incline us toward situational attributions for an action, towards seeing it not as self-disclosing. The psychological role of determinism is then to make such attributions readily available. I investigate this view experimentally in the context of very interesting and little-noticed feature of our thinking about responsibility and moral luck. But to begin, we turn to a philosophically attractive account of the person/situation distinction—the modal theory.
Chapter 1

The modal theory

1.1 Motivating a modal approach

One thing that stands out about John’s action is its extreme sensitivity to the details of his situation. Had John’s team lost the championship, or had the game been canceled, or had he dined at almost any other restaurant, or … John would have tipped the bare minimum.

This suggests an extremely plausible account of the distinction. Let’s say that an action is fragile to the degree that it would not have been performed in circumstances that depart in “certain ways” from the actual circumstances of its performance. Perhaps some actions would not have been performed if the wallpaper had been a shade lighter, or the temperature a degree cooler. Such actions would be highly fragile. The modal theory of the person/situation distinction identifies situationally caused actions with actions that are sufficiently fragile. So, to say that a kind action was situationally caused is to say that the agent would not have done the kind thing in relevantly similar circumstances; and to say that it was person-caused is to say that the agent would have continued to do the kind thing in a relatively broad range of circumstances (cf. Kelley, 1967).
My definition of *fragility* is purposefully vague. For *similarity* is a vague relation, so the question “Would S have done A in circumstances relevantly similar to C?” is also vague. The degree of fragility required for an action to count as situationally caused is therefore context-sensitive (cf., Lewis, 1979). So too is the question of which departures from the agent’s actual circumstances are relevant to whether her action is fragile. An agent who helps a stranger in her actual circumstances, but would not have helped had she not found a dime moments earlier, performs a highly fragile action; one who buys a stamp in her actual circumstances, but would not have done so had she not found a dime moments earlier, seems importantly different. This is presumably a consequence of our sense that whether one finds a dime should be irrelevant to whether one helps a stranger, but not to whether one buys a stamp.

Recent philosophical literature on situationism often appears to invoke a modal understanding of the person/situation distinction. John Doris, for example, writes,

> I allow for the possibility of temporally stable, situation-particular, ‘local’ traits that are associated with important individual differences in behavior. [T]hese local traits are likely to be extremely fine-grained . . . even seemingly inconsequential situational variations may ‘tap’ different dispositions (Doris, 2002, p. 25; also see Nelkin, 2005).

It is not surprising that philosophers would find a modal account of the person/situation distinction appealing: it is a straightforward extension of more general philosophical ideas about causation.

Lewis (1986) notes,

> When an effect depends counterfactually on a cause, in general it will depend on much else as well. If the cause had occurred but other
circumstances had been different, the effect would not have occurred. To the extent that this is so, the dependence is sensitive (p. 182).

Lewis’ example of highly sensitive causation is causing someone’s death by writing a letter of recommendation:

suppose I write a strong recommendation that lands someone a job; so someone else misses out on that job and takes another […] and there meets and marries someone; their offspring and all their descendants forevermore would never have lived at all, and a fortiori would never have died, and so presumably their deaths would not have occurred, but for my act (p. 184).

On Lewis’ analysis, counterfactual dependence is sufficient for causation, so it is proper to say that even highly sensitive counterfactual dependencies like this one are examples of causation. However, as Woodward (2006) notes, such examples often seem importantly different to examples of “real” causation. Highly sensitive counterfactual dependencies seem to us, at best, borderline cases of causation.

Given this, it is not surprising that we hesitate to cite John’s extremely fine-grained disposition to tip generously in just these circumstances as the cause of his action. Of course, his tipping does counterfactually depend on his internal dispositions, so one of the causes (in Lewis’ sense) of John’s generous action does lie within John himself. But we tend to ignore highly sensitive causes; hence our inclination to locate the cause of John’s action in the situation.

It may be objected that, in one sense, all actions are fragile: adding a hungry tiger or removing the oxygen from a room will extinguish most actions. Are two circumstances which are alike in all respects but for the addition of a hungry tiger similar enough to significantly influence our attributions? No; when assessing whether an action would have been performed in similar circumstances we are
ordinarily uninterested in how things would go in the presence of a hungry tiger or in the absence of oxygen. Can we specify a precise metric of similarity that would answer all such questions in advance? No; but the modal theorist might say that this is all to the good. For the question of whether an action is more or less person- or situation-caused is itself imprecise.

When defending his analysis of counterfactuals, Lewis considers that an overall similarity metric may be “hopelessly imprecise unless some definite respect of comparison has been specified.” He replies:

Imprecise though comparative similarity may be, we do judge the comparative similarity of complicated things like cities or people or philosophies—and we do it often without benefit of any definite respect of comparison stated in advance. We balance off various similarities and dissimilarities according to the importances we attach to various respects of comparison and according to the degrees of similarity in the various respects. Conversational context, of course, greatly affects our weighting of respects of comparison, and even in a fixed context we have plenty of latitude. Still, not anything goes. We have concordant mutual expectations, mutual expectations of expectations, etc., about the relative importances we will attach to respects of comparison. Often these are definite and accurate and firm enough to resolve the imprecision of comparative similarity to the point where we can converse without misunderstanding (Lewis, 1973, p. 420).

Perhaps this is the best reply to the objection that the modal theory is uninformative in lieu of a definite specification of the similarity relation. In any case, I will assume that counterfactual analyses are not hopeless, and I invite the reader to join me in this assumption. If the reader feels that such problems are insuperable, she will not
be tempted by a modal account of the person/situation distinction, and may safely skip this chapter.

1.1.1 Three classics of situationism

Situationist experiments provide excellent test cases for an analysis of the person/situation distinction. Since our aim is to understand the principles that lead people to explain actions in more or less agential or situational terms, our analysis should respect the verdicts of social scientists—for they are highly competent users of this folk theory—and, so far as possible, considered folk judgments. In the following section we will examine three classics of situationism, seeing in each case that the target actions are indeed modally fragile.

The Good Samaritan

Students at the Princeton Theological Seminary were told to prepare themselves to give a short sermon in a nearby building. In the hurried condition subjects were told: “You’re late; they were expecting you a few minutes ago; so you’d better hurry.” Subjects in the relaxed condition were told: “it will be a few minutes before they’re ready for you, but you might as well head on over.” On their way to the nearby building, subjects in both conditions encountered a victim “slumped in a door-way, head down, eyes closed, not moving.” As they approach, the victim coughs and groans, all the while keeping his head down. The question is: how many will stop to help?

While the general pattern of Darley and Batson’s (1973) result is not terribly surprising, the strength of the effect may be: 63 percent of the relaxed seminarians, but only 10 percent of the hurried seminarians, stopped to help the victim. Some of the subjects had been asked to give a sermon on the Good Samaritan, but they were no more likely to help than those giving a sermon on a topic that we would
not expect to focus their thoughts on kindness. Measured personality variables also failed to predict their behavior.

To see that the modal theory classifies the seminarians’ actions as situationally caused, consider one of the hurried seminarians who ignored the victim. What we learn from Darley and Batson is that if this hurried seminarian had been in the nearly identical circumstances of a relaxed seminarian, he probably would have helped the victim. Correspondingly, his action appears to be highly fragile, which, on the modal theory, is just to say that it was situationally caused, as it seems to be.

**Bystander apathy**

Are you more likely to get help in an emergency if you’re surrounded by a group of people or just one other person? While working on a questionnaire either alone or with a stooge, subjects in one of Darley and Latané’s (1968) studies heard the loud crash of someone taking a bad fall in an adjacent room, and then: “Oh, my God, my foot . . . , I . . . I . . . can’t move it. Oh . . . my ankle […] I . . . can’t get this . . . thing . . . off me.” While three quarters of lone subjects went to help the experimenter, only 7 percent did so in the presence of the unfazed stooge. The more people around, it seems, the less likely any one of them is to help.

In another study, while subjects filled out questionnaires either alone or with two confederates, a thick smoky substance would begin streaming into the room through an air vent. Unsurprisingly, the vast majority of lone subjects investigated the smoke and, after not too long, left the room to report it. In the second condition, the confederates would appear unfazed as the smoke streamed in. Now, in nine out of ten cases, the subjects remained “in the waiting room as it filled up with smoke, doggedly working on their questionnaire and waving the fumes away from their faces. They coughed, rubbed their eyes, and opened the window—but they did not report the smoke” (p. 218).
These findings are often cited as key examples of situationally caused behavior, and indeed this is the intuitive thing to say. The modal theory plausibly generates this result. Consider one of the lone subjects who reported the smoke to the experimenters. But for the small addition of an unmoved stooge, this subject almost certainly would not have reported the smoke. Thus her action appears to be highly fragile.

The Milgram experiment

You arrive at the Yale laboratory expecting to participate in an experiment on punishment and learning. The experiment requires a pair of subjects: one to play the role of learner, the other the role of teacher. You draw straws, and you get the role of teacher. (In fact, the learner is a confederate and the draw is rigged.)

The learner is to memorize word-pairs. He will be hooked up to an electric shock generator which you will control from an adjacent room. The generator’s control panel consists of 30 levers, labeled from 15 to 450 volts in 15-volt increments. You will read out the first word of a pair, followed by four possible answers. The learner will use an electronic signal to indicate his answer. At the first incorrect answer, you are to deliver the 15 volt shock. At each subsequent error you are to proceed to the next higher voltage.

At 75 volts, the learner begins to grunt in response to each shock. If you protest, the experimenter says, sternly but politely, “Please continue.” At 120 volts, the learner screams “Ugh! Hey this really hurts!” If you waver the experimenter will say: “The experiment requires that you continue, Teacher.” At 150 volts, the learner screams: “Let me out of here! You have no right to keep me here! My heart’s bothering me. Let me out!” If you look up for guidance, you’ll get a stony: “It is absolutely essential that you continue.” At 300 volts, the learner screams in agony: “I absolutely refuse to answer any more. Get me out!”
After you hit 330 volts, his cries and the sound of his fists pounding against the wall cease. He is silent for the rest of the experiment; presumably, he is comatose or dead. When you protest, you are told that “you have no choice,” you must treat his silence as an incorrect answer. The experiment continues like this until you either resist the experimenter’s entreaties four times without obeying, or depress the final, 450-volt, lever. Milgram’s famous result is that you are more likely than not to get to 450 volts. (For the full schedule of the learner’s protests, see Table 1.1 on the following page)

Recall Ross and Nisbett’s comment on Milgram’s result:

We do not find evidence that people are disposed to obey authority figures unquestioningly—even to the point of committing harmful and dangerous acts. [Milgram offers] a pointed reminder about the capacity of particular, relatively subtle situational forces to overcome people’s kinder dispositions (2011, p. 59, my emphasis).

Milgram’s experiment is often thought to pit two opposing forces against each other. On the one hand, there are the various features of the external situation: the Sorites-like sequence beginning at an innocuous voltage; your physical separation from the learner; the experimenter’s confident assurances; and so forth. On the other hand, there are facts about you: your belief that torture is morally wrong; your desire not to inflict horrible pain on another person; your concern and feelings of sympathy for the learner. What we are supposed to learn from Milgram is that in circumstances like these, the situation wins out nearly every time.

The modal theory helps to explicate these metaphors. We believe that most of these subjects are ordinary, morally decent people. Thus, although we may not be able to disaggregate the features of the situation, we’re confident that but for some feature of the experimental setup, these subjects would not have obeyed the
Table 1.1: The learner’s schedule of protests in Milgram’s obedience studies.

<table>
<thead>
<tr>
<th>Voltage</th>
<th>Audible protest from the learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>75 volts</td>
<td>Ugh!</td>
</tr>
<tr>
<td>90 volts</td>
<td>Ugh!</td>
</tr>
<tr>
<td>105 volts</td>
<td>Ugh! (louder)</td>
</tr>
<tr>
<td>120 volts</td>
<td>Ugh! Hey this really hurts.</td>
</tr>
<tr>
<td>135 volts</td>
<td>Ugh!</td>
</tr>
<tr>
<td>150 volts</td>
<td>Ugh! 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.</td>
</tr>
<tr>
<td>165 volts</td>
<td>Ugh! Let me out! (shouting)</td>
</tr>
<tr>
<td>180 volts</td>
<td>Ugh! I can’t stand the pain. Let me out of here! (shouting)</td>
</tr>
<tr>
<td>195 volts</td>
<td>Ugh! Let me out of here. Let me out of here. My heart’s bothering me. Let me out of here! You have no right to keep me here! Let me out! Let me out! Let me out! Let me out of here! My heart’s bothering me. Let me out! Let me out!</td>
</tr>
<tr>
<td>210 volts</td>
<td>Ugh! Experimenter! Get me out of here. I’ve had enough. I won’t be in the experiment any more.</td>
</tr>
<tr>
<td>225 volts</td>
<td>Ugh!</td>
</tr>
<tr>
<td>240 volts</td>
<td>Ugh!</td>
</tr>
<tr>
<td>255 volts</td>
<td>Ugh! Get me out of here.</td>
</tr>
<tr>
<td>270 volts</td>
<td>(Agonized scream.) Let me out of here. Let me out of here.</td>
</tr>
<tr>
<td>285 volts</td>
<td>(Agonized scream.)</td>
</tr>
<tr>
<td>300 volts</td>
<td>(Agonized scream.) I absolutely refuse to answer any more. Get me out of here. You can’t hold me here. Get me out. Get me out of here.</td>
</tr>
<tr>
<td>315 volts</td>
<td>(Intensely agonized scream.) I told you I refuse to answer. I’m no longer part of this experiment.</td>
</tr>
<tr>
<td>330 volts</td>
<td>(Intense and prolonged agonized scream.) Let me out of here. Let me out of here. My heart’s bothering me. Let me out, I tell you. (Hysterically) Let me out of here. Let me out of here. You have no right to hold me here. Let me out! Let me out! Let me out! Let me out! Let me out! Let me out!</td>
</tr>
</tbody>
</table>
experimenter’s instructions. That is to say, we’re confident that their behavior is modally fragile.

The results of Milgram’s variations on his initial experiment strengthen this impression. Milgram found that: subjects who participate in an office building in Bridgeport, Connecticut, rather than in a Yale laboratory, are less likely to obey; subjects who receive the experimenter’s instructions over the phone, rather than in person, are much less likely to obey; subjects in identical circumstances but for the addition of two disobedient fellow teachers, almost certainly will not obey; and so on. All this suggests that subjects’ obedience is indeed fragile. Hence, the modal theory generates the intuitive verdict that it is situationally controlled.

1.1.2 Situational causation and self-disclosure

A person who acts because of a compulsive psychological disorder or because she is at the limits of enervation, does what she wants to do. There’s often no doubt about it, the addict wants to shoot up, and the kleptomaniac wants those shiny things in her pocket. How else could we explain their behavior? However, something seems awry in their psychology; their actions can seem disconnected from who they truly are deep down.

The same is often true of subjects in the famous situationist experiments. Here is an observer’s report on the Milgram experiment:

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.” And yet he continued to respond to every word of the experimenter, and obeyed to the end (1963, p. 377).

Almost all subjects in Milgram’s original experiment experienced extreme anxiety: they 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” (ibid., p. 375; see Table 1.2 on the next page).

With this in mind, ask yourself: when Milgram’s subjects deliver excruciating electric shocks to a desperate victim who has explicitly revoked consent, do their actions seem to reveal who they truly are deep down? I think the intuitive thing to say is that we do not see their true selves reflected in their obedient behavior; if anything, we see it in their sweating, trembling, and lip biting. Intuitively, their obedience is not self-disclosing.

Common philosophical analyses are able to capture this result without much difficulty. Take the Frankfurtian theorist. She holds that actions are self-disclosing just in case they are caused by desires that an actor wants to lead her to action (Frankfurt, 1971). Milgram’s subjects plausibly lack second-order desires of this sort—they do not want their desire to obey (or to avoid decisively confronting the experimenter) to effectively lead them to action. So the Frankfurtian theorist can account for the intuitive verdict. The same goes for the Watsonian theorist. She holds that an action is self-disclosing just in case the desires which motivate it are sourced in the actor’s values (Watson, 1996); and it is very implausible that the desires which motivate most of Milgram’s subjects are sourced in their values. Or take the rationalist theorist. She holds that an action is self-disclosing just in case it reflects an actor’s judgment-sensitive attitudes, i.e., the attitudes that would exist in an ideally rational person only when she judges that there are sufficient reasons for them (Smith, 2005; Scanlon, 1998). It seems she too might build a plausible case for the intuitive verdict. And the same goes for many other classic situationist results:
PROZI: I can’t stand it. I’m not going to kill that man in there. You hear him hollering?
EXPERIMENTER: As I told you before, the shocks may be painful, but—
PROZI: But he’s hollering. He can’t stand it. What’s going to happen to him?
EXPERIMENTER (his voice is patient, matter-of-fact): The experiment requires that you continue, Teacher.
PROZI: Aah, but, uhh, I’m not going to get that man sick in there—know what I mean?
EXPERIMENTER: Whether the learner likes it or not, we must go on, through all the pairs.
PROZI: I refuse to take the responsibility. He’s in there hollering!
EXPERIMENTER: It’s absolutely essential that you continue, Prozi.
PROZI: There’s too many left here, I mean, Jeez, if he gets them wrong, there’s too many of them left. I mean, who’s going to take the responsibility if anything happens to that gentleman?
EXPERIMENTER: I’m responsible for anything that happens to him. Continue, please.
PROZI: All right. (Consults list of words.) The next one’s “Slow—walk, truck, dance, music.” Answer, please. (A buzzing sound indicates the learner has signaled his answer.) Wrong. A hundred and ninety-five volts. “Dance.” (Zumph!) LEARNER (yelling): Let me out of here! My heart’s bothering me! (Teacher looks at experimenter.)
PROZI: I don’t think he is going to answer. EXPERIMENTER: If the learner doesn’t answer in a reasonable time, about four or five seconds, consider the answer wrong. And follow the same procedures you have been doing for wrong answers. Say, “Wrong,” tell him the number of volts, give him the punishment, read him the correct answer. Continue, please.
PROZI: “Green—grass, hat, ink, apple.” Answer, please. (More silence. Subject just sits there.) EXPERIMENTER: Go on with the procedure, please.
PROZI: 315 volts. The answer is “ink.” (Zumph!) LEARNER: Ohh. Ooohh.
PROZI (relieved to hear response): Next one.
EXPERIMENTER: I absolutely refuse to answer. PROZI: You got to. You get a shock if you don’t answer. (Experiment continues with learner refusing to answer, and, finally, failing even to grunt or groan. The subject stands up.)
PROZI: I think something’s happened to that fellow in there. I don’t get no answer. He was hollering at less voltage. Can’t you check in and see if he’s all right, please?
EXPERIMENTER (with same detached calm): Not once we’ve started. Continue, Teacher.
PROZI (sits down, sighs deeply): “Cool—day, shade, water, paint.” Answer, please. Are you all right in there? Are you all right?
EXPERIMENTER: Please continue, Teacher. Continue, please. (Subject pushes lever. Zumph!)
PROZI (swiveling around in his chair): Something’s happened to that man in there. (Swiveling back.) “Low—dollar, necklace, moon, paint.” (Turning around again.) Something’s happened to that man in there. You better check in on him, sir. He won’t answer or nothing.

EXPERIMENTER: Continue. Go on, please.

PROZI: You accept all responsibility?

EXPERIMENTER: The responsibility is mine. Correct. Please go on. (Subject returns to his list, starts running through words as rapidly as he can read them, works through to 450 volts.)

PROZI: That’s that.

EXPERIMENTER: Continue using the 450 switch for each wrong answer.

PROZI: But I don’t get anything!

EXPERIMENTER: Please continue. The next word is “white.”

PROZI: Don’t you think you should look in on him, please?

EXPERIMENTER: Not once we’ve started the experiment.

PROZI: What if he’s dead in there? (Gestures towards the room with the electric chair.) I mean, he told me he can’t stand the shock, sir. I don’t mean to be rude, but I think you should look in on him. All you have to do is look in on him. All you have to do is look in the door. I don’t get no answer, no noise. Something might have happened to the gentleman in there, sir.

EXPERIMENTER: We must continue. Go on, please.

PROZI: You mean keep giving him what? Four-hundred-fifty volts, what he’s got now?

EXPERIMENTER: That’s correct. Continue. The next word is “white.”

PROZI (now at a furious pace): “White—cloud, horse, rock, house.” Answer, please. The answer is “horse.” Four hundred and fifty volts. (Zzumph!) Next word, “Bag—paint, music, clown, girl.” The answer is ”paint.” Four hundred and fifty volts. (Zzumph!) Next word is “Short—sentence, movie . . .

EXPERIMENTER: Excuse me, Teacher. We’ll have to discontinue the experiment.

The common philosophical accounts can all plausibly capture the intuitive judgment that subjects’ actions are not self-disclosing.

The modal theory, in concert with any of these analyses, helps to explain why. Consider an analysis of self-disclosure according to which an action is self-disclosing if and only if it flows from desires backed up by second-order volitions (i.e., desires that an agent not only has, but wants to act on). On this analysis, it makes sense that actions which are not self-disclosing are more modally fragile than those which are self-disclosing: an alienated agent who desires something but wishes that she did not is more liable to waver in her pursuit of that thing than an agent who not only desires it, but is glad to do so. The same seems to go for actions which are sourced in an agent’s values and those which reflect her judgment-sensitive attitudes.¹

¹This is not intended an exceptionless generalization—the kleptomaniac is a good example for making that point. What I am trying to explain is why actions that seem to be situationally caused rarely seem to be self-disclosing. In other words, I am trying to explain why our judgments about these two things co-travel. To explain this, it is enough if the following is a bit of folk theory: typically, a person who reflectively endorses what she is doing is less likely to waver than she would
It appears, then, that the modal theory provides a good analysis of the person/situation distinction, one which grows out of an established philosophical approach to causality, and helps to explain why situationally caused actions are rarely, if ever, self-disclosing. There’s only one problem: the modal theory is false.

1.2 Counterexamples to the modal theory

Let’s return to the Good Samaritan experiment. Recall that students at the Princeton Theological Seminary were told to prepare themselves to give a short sermon in a nearby building. In the hurried condition subjects were told: “You’re late; they were expecting you a few minutes ago; so you’d better hurry.” Subjects in the relaxed condition were told: “It will be a few minutes before they’re ready for you, but you might as well head on over.” On their way to the nearby building, subjects in both conditions encountered a victim “slumped in a door-way, head down, eyes closed, not moving.” While most of the relaxed seminarians stopped, only 10% of the hurried ones offered to help.

Let’s focus on Helpful, one of the seminarians in the relaxed condition who helped the victim. Helpful is on his way to the lecture theater, thinking about how he will begin his talk, when he sees the distressed victim in the doorway. Helpful stops and asks the victim if he is alright. Now, does it seem more appropriate to say that Helpful’s kind action is caused by the situation in which he finds himself, or does it seem that the action arises from deep within Helpful himself? Even knowing what Helpful would probably have done had he been randomized into the rushed condition, it still seems clear that he himself is the cause of his kind action.²

²While I attempt to test cases that seem to me potentially unclear, sometimes my arguments rely on my own armchair intuitions and my expectation that the reader will share them. There are no principled differences between the cases where I present evidence and those where I merely be were she to reflectively repudiated it. An ordinary way of saying this is: reflectively conflicted people are flaky.
Now, let’s focus on Hurried, one of the seminarians in the rushed condition who ignored the victim. Hurried is walking briskly to the lecture theater, he sees the victim, and perhaps he registers somewhat dimly that the victim could need help. *But he doesn’t stop; he just walks on.* In this case, it seems clear that Hurried’s callous action is, by comparison with Helpful’s kind action, caused by the situation. Thus, an interesting pattern emerges. While we are more inclined to attribute Hurried’s unhelpful action to the situation, we are more inclined to attribute Helpful’s action to the person. Why should this be, and what can the modal theorist say about it?

Recall that the modal theory holds that we become inclined to judge that an action is caused by the situation to the degree that we judge that it would not have been performed in circumstances relevantly similar to the actual circumstances of its performance. (The word *relevantly* indicates that the similarity relation is vague and can be assessed only against a background of common knowledge and conversational context.) This modal analysis appears to get things right in Hurried’s case. For we know that Hurried could *easily* have found himself in Relaxed’s circumstances, and had he found himself there, he too (we may suppose) would have helped the victim. In relevantly similar circumstances Hurried would not have performed his callous action; so the modal theory implies that his action is caused by the situation, just as it seems to be.

But just as Hurried might have helped had he been in the same circumstances as Relaxed, so Relaxed might *not* have helped had he been in the same circumstances as Hurried. There is no reason to think the cases *modally* asymmetrical—each could have found himself just as easily in the other’s circumstances. But even when we report my own reactions—all are intended to shed light on a folk theory. So, ideally, all claims about what seems intuitive would be backed up by the results of rigorous experimentation. But running experiments is costly and we philosophers are members of the folk: our own judgments (especially the clearest of them) are often a good guide to how many other people will respond to our cases. That’s why the results of experimental philosophy studies are, often enough, unsurprising. (And, as replication studies sometimes suggest (e.g., Kim and Yuan, 2014; Seyedsayamdost 2014, 2015) when the results are surprising this can be because the studies were underpowered or otherwise methodologically flawed, and not because philosophers’ judgments were unrepresentative.)
hold this fact firmly in mind, I claim that one action still seems more situation-caused than the other. Since we believe their actions are equally fragile, the modal theory cannot explain why our judgments about them should differ.

My judgments about Hurried and Relaxed are clear, but they are contrary to the way social psychologists speak, and since the question we are addressing is whether ordinary people have something like the modal theory in their heads, it would be good to see if results like these turn up in a more diverse pool of participants. Perhaps the easiest way to see that the modal theory does not capture the person/situation distinction is to test a case like the following:

Jones grew up in the Australian outback and has always been fascinated by the natural world. Consequently, she studied environmental engineering and now finds herself working as an environmental supervisor at a large oil company. Her job is to evaluate the likely environmental impact of drilling at promising sites.

Jones knows that the company’s most recent plan risks seriously harming a delicate marine ecosystem, but the company has been struggling financially and is thus pressuring Jones to approve the plan. Her superiors have even intimated that if Jones won’t do it, they’ll find someone else to take her job.

When Jones’ manager brings the paperwork for her to sign, she thinks to herself: “If I don’t do it, someone else will,” and she picks up her pen to sign the approval documents. But just as Jones is about to sign the documents a striking and very beautiful bird flies past her window, catching her eye. It is the very same kind of bird that lived by her childhood home. Jones puts down her pen and refuses to sign. If it comes to it, she will resign from the company.
When presented with this case, a large majority of people respond that Jones’ action arises more from within herself—she refused to sign mainly because of her deep respect for the natural world. That’s the dispositional attribution. Most people do not judge that Jones’ action arises more from the situation in which she found herself. The situational attribution—that she refused because the bird caught her attention—seems to get the focus all wrong. But notice that Jones’ refusal is indeed modally fragile: if the bird had appeared a moment later it wouldn’t have caught her eye; or if it had been some other kind of bird she wouldn’t have been reminded of her childhood; or if . . .

In the most similar possible worlds Jones does approve the plan. And yet in the actual world, when she refuses to approve the plan, her behavior does not seem to be situationally controlled. To the contrary, it seems to reflect something about who Jones is at her very essence. So the modal theory has got the answer wrong. Even though her action closely depends on the very finest details of her situation, the situation isn’t what seems most important to explaining it.

---

3North American subjects were recruited on Mechanical Turk (38% female, mean age 38 years, N = 100). After reading the vignette above, they were asked “Why did Jones refuse to sign the approval documents?” Response alternatives were: “Mainly because she noticed the bird” and “Mainly because she has a deep respect for nature,” and were counterbalanced for order. 79% of respondents selected the person attribution. To check that participants were convinced that Jones’ action was modally fragile I asked what would have happened if Jones had not noticed the bird. 86% responded that “Jones probably would have signed the approval.” Subjects who responded “Jones probably would still have refused to sign” were excluded.

Because this clear result might be an effect of forcing participants to choose between dichotomous response alternatives (Cullen, 2010), I tested the same vignette on a different sample using Likert scales (37% female, mean age 35 years (SD = 12.8), N = 103). When the same question was posed with a 9-point Likert scale (1 = mainly because of her love for the natural world; 9 = mainly because of the bird) the mean response was 3.3 [2.77, 3.83] (SD = 2.3). The mean response to the question, “Does the cause of Jones’ refusal to sign lie more within Jones herself, or more within the situation Jones is in?” (1 = within Jones herself; 9 = in the situation), was 2.9 [2.39, 3.41] (SD = 2.2). The questions were counterbalanced for order. I found no significant effect of age or sex in either sample.
1.2.1 Amending the modal theory

Because it cannot capture our intuitions, the simple modal theory outlined above does not provide a psychologically adequate analysis of the person/situation distinction. But perhaps an amended modal theory might do better. Consider the following case (I owe it to Johann Frick):

A competitive sharpshooter is lining up a challenging shot, but just as she prepares to squeeze the trigger, a spectator’s watch reflects a beam of sunlight into her eyes. She would have missed the target were it not for a stray golf ball flying by, blocking the sunlight and giving her just enough time to refocus and make the shot.

We may suppose that in nearby possible worlds the golf ball fails to block the reflection, and Sharpshooter misses the shot. So her action—making the shot—is modally fragile. Yet it does not seem correct to attribute her success to the situation.

Perhaps the modal theory only generates the implausible verdict that Sharpshooter’s success is attributable to her situation because it restricts our consideration to the very nearest possible worlds. If we suppose that the nearest worlds do not include the perfectly placed golf ball, then consideration of just those worlds will make it seem that Sharpshooter’s actual success is somehow out of the ordinary. But sharpshooter occupies a most unusual region of logical space, and while it may be true that in the nearest possible worlds she misses the shot, in vastly more worlds in which she takes a shot of similar difficulty, she has a clear view of her target. And in these worlds she makes the shot.

Sharpshooter’s case is analogous to Jones’. The role of the sunlight is played by Jones’ boss pressuring her to approve the plan, thereby masking her environmentalism. The role of the golf ball is played by the bird that reminds Jones of her childhood, thereby masking the masking effect of her boss. So perhaps we should
say the same thing about Jones’ case: her refusal to sign the approval documents is mostly attributable to her, and not to the situation, because in most possible worlds in which she must choose whether to approve a dangerous drilling plan, she is not under undue pressure, and in these worlds she’s doing just what we expect her to do—rejecting dangerous plans (see Figure 1.1).

Thus the amended modal theory holds that

an action is person- (or situation-) caused to the extent that the agent does (or does not) perform the action in sufficiently many worlds in which she faces a choice of the relevant type.

While comparative similarity still determines the relevant set of worlds at which the action must occur, the amended modal theory does not impose a similarity ordering on this set of nearby worlds. This means its verdicts are not swayed because the action happens to take place in an unusual region of logical space—for example, in a region in which the agent is under pressure to approve a dangerous plan. Rather, it asks us to assess what the agent would do in all relevantly similar circumstances.

**Figure 1.1:** Comparison of the worlds considered by the simple and amended modal theories.
in which she faces a choice of the relevant kind. (The amended modal theory is in this respect somewhat akin to the strict conditional analysis of counterfactuals.)

This view gets many cases right. Think back to Stingy John. It is certainly true that in most worlds in which John must choose how much to tip, he chooses to leave the bare minimum. And it seems believable that even the seminarians who ignore the victim in Darley and Batson’s (1973) experiment, decide to help people in most worlds in which they face a relevantly similar choice. Similarly, in most worlds in which Milgram’s subjects must decide whether to torture an innocent victim, they of course choose not to.

Unfortunately, the amended modal theory also faces counterexamples. Consider the following case:

Ever since he was a small child, Mort has felt, deep down inside, that there was something calling him to be a woman. Unhappily, however, he lives in a bigoted and unforgiving age, so he almost never acts on these feelings.

One day, Mort goes to see Marilyn Monroe star in the film Some Like It Hot. Even though this would not usually have any effect on his behavior, after watching the film Mort feels a strong desire to express his transgendered identity. That night, once the house is very quiet and everyone is asleep, he puts on his wife’s makeup and gets dressed in one of her evening gowns.

Mort’s action marks a vanishingly rare spec in his region of logical space: in almost all circumstances in which he must choose whether to express his transgendered identity, he behaves in the way his society demands. But to many people, the cause of Mort’s action will seem to lie deep within Mort himself, so the amended modal theory cannot account for our judgments.
It may be objected: sure, Mort wouldn’t get into his wife’s dress after watching *On the Waterfront*, but that’s not relevant. The relevant comparison class includes only those worlds in which he sees a film like *Some Like It Hot*—a film in which the male leads disguise themselves as women after witnessing a mob hit—and maybe we believe that in many of those worlds Mort will express his transgendered identity (despite the vignette asserting otherwise). So the objector holds that this putative counterexample requires us to consider an artificially wide variety of counterfactual circumstances. The contextually determined class of relevant possibilities might really include only those worlds in which Mort has been placed in *very* difficult circumstances—circumstances which mask his disposition to conform to his society’s expectations.

I reply that even if we stipulate that Mort gives into his feelings in only ten percent of the worlds in which he sees *Some Like It Hot*, when he dresses in his wife’s gown and puts on her makeup, his actions still will not seem to arise from the situation in which he finds himself. Even if people explicitly imagine that seeing *Some Like It Hot* would almost invariably leave Mort cold, this will not have any effect on their judgments. I conclude that cases like this one count strongly against the amended modal theory. But the reader may not be convinced. Never mind—in Chapter 2 we will see experimental data that suggests the pattern of responses to cases like Mort’s is actually far more complicated than even the amended modal theory predicts.

### 1.2.2 Individual differences

I believe that the winner of the Tour de France could easily have failed to win—an enormous number of conditions had to be in place for him to win as he did. But this does not incline me to think that winning the Tour is situation-caused. The same is true of other unlikely achievements: becoming a great musician or artist or...
revolutionary is always a modally precarious matter. Yet these achievements often seem self-disclosing and their causes seem to lie firmly within the agents themselves. Similarly, the counterexamples we have considered show that our judgments are not always driven by whether an action appears modally fragile—Jones’ action is exquisitely sensitive to the finest details of her situation, but it does not seem to be caused by the situation.

Such cases refute the simple modal analysis but might yet be explained in somewhat different modal terms: perhaps we draw on our beliefs about whether most other agents would have performed the action if they had been in the actual agent’s circumstances. Presumably we believe that no matter the circumstances, most people will not win the Tour or become a great musician—most of us are simply not physically or mentally capable of such feats. So maybe this is what the person/situation distinction tracks. That is, perhaps situational attributions are, as Ross held, those that do not “imply any dispositions on the part of the actor beyond those typical of actors in general,” and person attributions are those that “imply unique, relatively atypical or distinguishing personal dispositions” (Ross 1977, p. 177, my emphasis).

Although this “individual differences” analysis generates the somewhat odd consequence that whether an action is person-caused or situation-caused does not supervene on facts about that person and her environment, it certainly gets many of the problem cases right. Even in Jones’ exact circumstances most people, even most people who have a deep love of the natural world, would acquiesce; they would sign the approval documents. Jones’ disposition to respond to a certain kind of bird by refusing to sign is indeed highly distinguishing.

But this cannot be the whole explanation for why we judge that some actions are person-caused. For we believe that most people in Helpful’s circumstances would indeed help the victim—his action arises from a disposition “typical of actors in
general”—and yet it appears to be person-caused. So some person-caused actions arise from common dispositions. (If you doubt that most people would help an adult stranger, imagine instead that the victim is a child or a close friend.)

Similarly, some situation-caused actions arise from highly unique dispositions. Consider Little Albert:

When he was a baby, Little Albert was the subject of a psychology experiment in which he was conditioned to fear furry white things. This conditioning was extremely successful, producing in Albert a stable disposition to react to furry white things with violent fear. One day, when Albert is all grown up, he sees a man with an extremely furry white beard. Instantly filled with fear, Albert punches the man in the nose and flees from the scene.

Even though the explanation for Albert’s action implies that he has a highly unique personal disposition, compared to many other people’s nose-punchings, his action does not seem highly person-caused. So distinguishing dispositions can undergird more situation-caused actions like Albert’s, and common dispositions can undergird more person-caused actions like Helpful’s. So the individual differences analysis cannot be right.

It may be objected that although we are not solely interested in whether most other agents would have performed the action, we may balance individual differences against modal fragility, weighing covariation with the situation more heavily in some cases and covariation with the person more heavily in others (e.g., Pruitt and Insko, 1980; Kelley, 1973). In cases like Albert’s we give little weight to whether most other people would perform the action—in these cases we are interested in modal fragility, in what Albert would do were he in relevantly similar circumstances,
for example, those in which the man is clean-shaven. But in other cases, like Jones’, we might give individual differences greater weight.

We do not need to consider how this weighting might work to see that it cannot provide a psychologically adequate account of the person/situation distinction. For we can imagine person-caused actions that most people would perform, but only in very specific circumstances; that is to say, some person-caused actions are both modally fragile and caused by widely shared dispositions. Indeed, Helpful’s checking to see if the victim is all right is a good example. His action is both modally fragile—in relevantly similar circumstances he would have ignored the victim—and widely believed to be caused by very common dispositions—most people in his circumstances would indeed offer to help the victim. And yet the cause of his action appears to lie within Helpful himself. Similarly, some situation-caused actions arise from highly distinguishing, modally robust dispositions. So there is no weighting of modal fragility and individual differences that could generate our judgments.

1.3 Conclusion

The modal theory grows out of an established approach to understanding causation, it helps to explain why situationally caused actions are rarely, if ever, self-disclosing, and it helps to operationalize the ways that many social psychologists have thought about the person/situation distinction. But as the cases considered in this chapter show, it does not provide an adequate account of the principles that people actually rely on when attributing actions. These cases do not rely for their force on locating agents in unusual regions of logical space, or on ignoring how an action might covary across different agents in the same circumstances.

Counterexamples like theses strongly suggest that the modal theory is false, but they do not help us to see why it is false or, more pertinently, to understand
what’s actually going on when people locate the cause of an action in more or less situational factors. In the next chapter I argue for an answer to this question that is completely unlike those we have considered so far.
Chapter 2

The mismatch theory

On the modal theory, beliefs about how agents would act in relevant counterfactual circumstances drive our self-disclosure judgments and person/situation attributions. One model compatible with this view holds that judgments of self-disclosure and attribution are at best covariates: an action-theoretic judgment might (1) generate judgments about whether an action is self-disclosing and (2) also inform judgments about how an agent would act in various counterfactual circumstances (see Figure 2.1). These counterfactual judgments in turn could then generate person or situation attributions, as the case may be.

According to the theory I prefer—the mismatch theory—the connection is much

Figure 2.1: A plausible model of the relationship between self-disclosure judgments and person/situation attributions, compatible with the modal theory.
closer. Indeed, on my view, the simplest and best explanation of why our attribut-
tional and self-disclosure judgments so reliably co-travel, and why situationists
so often appear to invoke tacitly the folk psychological notion of the true self, is
that the principles that we rely on to determine whether an action is self-disclosing
are the very same principles that drive us toward person or situation attributions.1
Thus, on the mismatch theory, moral judgment is at the root of both attribution
and self-disclosure: the boundary between the person and the situation is drawn in
moral ink and the relation of self-disclosure can only be characterized in evaluative
terms (see Figure 2.2 on the following page). The mismatch theory thus joins a
growing body of evidence for the pervasive role that moral attitudes play in human
cognition—even when its object is not obviously moral (Knobe, 2010).

1Shamik Dasgupta suggested the following as a case where self-disclosure and person-attribution
seem to come apart:

Tony has devoted all his life to humanitarian work. He toils endlessly, 80 hour weeks,
for years, at great personal loss, etc. But to keep up his good work, Tony occasionally
needs to unwind. The most effective way for him to unwind is to engage in [insert
morally objectionable activity]. Knowing this, Tony plans a weekend away every year,
during which he engages in this activity.

The idea is that, while Tony’s engaging in this bad activity does not seem self-disclosing, the cause
does appear to lie within Tony himself. My own judgments about these cases are unclear, and I have
not found a version where people’s responses come apart in the predicted way. For example, when
the objectionable activity is cheating on his wife, people judge that Tony’s action is both highly
self-disclosing and that its cause lies almost entirely within Tony himself. When I tested this vignette
on mTurk (N = 50; M_\text{age} = 34.4 (SD = 11.1); 48% female) the average response on a 9-point Likert scale
to both the self-disclosure question (mean = 7.1 [6.51, 7.74] (SD = 2.2)) and the attribution question
(mean = 8.3 [7.99, 8.63] (SD = 1.13)) were extremely high. I have never seen such unambiguous
person-attributions for a morally bad action, suggesting that participants are not at all convinced
that Tony is a good person. And indeed, even when the vignette emphasizes that he has a morally
good essence, and that he views his cheating as a deeply regrettable necessity, remarkably, 70% of
participants respond that he is, at heart, morally bad. This pattern holds for both male and female
participants. Very few things, it turns out, are thought to be more self-disclosing than infidelity.

When Tony’s weekend away involves something generally regarded as less objectionable than
cheating on his wife (for example, hunting deer) people tend to accept that he is morally good, but
then both self-disclosure judgments and person-attributions tend to be reduced.

Still, people often have explicit beliefs about the true self, and it seems reasonable to suppose
that in the right mindset these beliefs might influence their self-disclosure judgments without
strongly influencing their attributions. I think the mismatch theory should therefore be understood
as attempting to describe just one very important determinant of both self-disclosure judgments and
person/situation attributions.
Here are the details. The mismatch theory holds that we become inclined toward situational attributions when there is a mismatch between the perceived moral status of an action and the perceived moral status of the actor. We express this mismatch when we say the action is caused or explained not by the person, but by the situation. Since two factors are needed to get a mismatch, there are two ways for an action to be situationally controlled. The theory can be stated as follows:

an observer will be inclined to judge that an action is situationally caused to the extent that either (a) she judges that the action is morally bad and that the actor is morally good, or (b) she judges that the action is morally good and that the actor is morally bad.

To take our previous example, consider again the surprising acquiescence of Milgram’s subjects. We are inclined to describe their obedient behavior as caused or explained by the situation. Why? The mismatch theory answers: because we judge that torturing innocent strangers at the polite insistence of a psychologist is morally bad, and we believe that the participants are, on the whole, morally good people.

As stated the mismatch theory leaves a lot unexplained. What is it about some explanations that leads us to think of them as ‘situational’? What are we judging when we judge that an agent is morally good or bad? Since there are several plausible interpretations of the mismatch theory, and since the considerations that
follow tend to support the theory on any plausible interpretation, it makes sense to separate the evidence for the more ‘formal’ structure of the theory from its substantive interpretation. We return to consider substantive interpretations of the theory in Section 2.4.

2.1 Extant evidence for the mismatch theory

Apart from capturing many of our intuitions about self-disclosure and morally valenced examples of the person/situation distinction, some of the best evidence for the mismatch theory comes from its ability to explain intergroup attributional patterns as well as a very striking feature of the actor-observer literature.

2.1.1 The Actor-Observer Asymmetry

Jones and Nisbett (1972) famously hypothesized that “actors tend to attribute the causes of their behavior to stimuli inherent in the situation, while observers tend to attribute behavior to stable dispositions of the actor” (p. 93). These two tendencies roughly correspond to the fundamental attribution error (someone cuts you off on the way home from work—he’s obviously a jerk) and self-serving bias (you cut someone off—you’re just in a hurry to pick up your daughter).

How does the mismatch theory explain this asymmetry? It is well-known that we are generally more confident of our own moral goodness than we are of others’. People tend “to think they are more charitable, cooperative, considerate, fair, kind, loyal, and sincere than the typical person but less belligerent, deceitful, gullible, lazy, impolite, mean, and unethical—just to name a few (Alicke, 1985; Allison et al., 1989; Dunning et al., 1989; Goethals et al., 1991)” (Epley and Dunning, 2000, p. 861). Given this, the mismatch theory provides a natural explanation for the actor-observer bias in the case of morally bad behaviors. For any given bad behavior
is more likely to conflict with one’s self-assessment than with one’s assessment of another. That’s why we are all more likely, relative to base rates, to view our own blameworthy acts as situationally caused.\(^2\)

But what about good actions? As it’s described by social scientists, the actor-observer asymmetry applies equally to morally good and bad behavior: it actually predicts that people prefer situational attributions for their own behavior, good or bad. If this were true—if there really were a morally neutral, actor-observer bias—the mismatch theory would be in serious trouble. Thankfully, a surprising recent meta-analysis of the enormous actor-observer literature shows that there is no such bias.\(^3\)

What the literature does show is that we are, relative to base rates, more likely to provide situational attributions for our own actions when they seem blameworthy

\(^2\)This explanation may seem to be in tension with my explanations for people’s responses to the cases discussed in Chapter 1 (and many of those to come in this and the following chapter). To explain, for example, why the cause of Jones’ action seems to lie within Jones herself, I assumed that we believe Jones is morally good. In all the other cases, too, I relied on the assumption that we default to the belief that the agents are morally good. (Although this assumption seems to run counter to recent and long-established findings on impression formation (Mende-Siedlecki et al., 2013; Skowronski and Carlston, 1989), we will continue to encounter evidence for an important kind of positivity bias throughout this dissertation. See § 3.2.3 on page 90 for a brief discussion.) But now, if we believe that most people are morally good, then it seems we should not be more likely to give person-attributions for other people’s blameworthy actions. It might seem, to the contrary, that we should actually be equally likely to give situation-attributions for their blameworthy actions. In other words, we should exhibit the same biases when explaining others’ actions as we do when explaining our own, and this appears to directly contradict the explanation just given for why we are all more likely, relative to base rates, to view our own blameworthy acts as situationally caused.

I think this tension is for the most part merely apparent. The first thing to note is that the effect sizes for actor-observer asymmetries are typically small, and smaller still when people are explaining positive actions. The largest effects are found among Western participants in between subjects studies. For such participants, \(d = 0.36\) on average for negative actions, and \(d = 0.24\) on average for positive actions (\(\eta^2 = 20\%\)) (Malle, 2006, p. 903). The fact that the asymmetry tends to be smaller when explaining positive actions suggests that positive actions conflict less strongly with both self and other evaluations. So to explain the actual patterning of results we do not need to give up on the assumption that we believe most people are morally good; we only need, in addition, the well-evidenced assumption that we often believe ourselves to be somewhat better. The greater the disparity between one’s self-evaluation and one’s evaluation of others, the greater the (valenced) asymmetries will be, and this will hold true whether self and other are thought to be on the same or opposite sides of “morally neutral.”

\(^3\)Malle’s (2006) meta-analysis found that a “total of 61 studies had actors and observers explain negative events, and for those, the classic asymmetry pattern was obtained on average. Contrary to that, in 45 studies, actors and observers explained positive events and displayed a reverse asymmetry” (p. 902).
and dispositional attributions when they seem praiseworthy. The pattern actually reverses when we judge the actions of other people: we are more likely, relative to base rates, to provide dispositional attributions for other people’s actions when they seem blameworthy and situational attributions when they seem praiseworthy (see Figure 2.1 on the following page). Given that we tend to think we’re morally better than average, the broad patterning of the actor-observer literature revealed in meta-analysis is exactly as the mismatch theory predicts. This constitutes strong support for the theory. On the present account, the fundamental attribution error and self-serving bias are really just opposite sides of a single coin: they are both a product of our expectation that people and their behavior will morally match up, and our belief in our own moral superiority.

### 2.1.2 Intergroup attributions

We have so far considered how one’s moral evaluation of oneself interacts with one’s moral evaluation of some arbitrary other agent. When we consider self-identified groups of agents the mismatch account leads to the following prediction: insofar as in-groups think of themselves as morally better than out-groups, in-group members will be more likely, compared to base rates, to produce dispositional attributions for their own members’ praiseworthy acts and situational attributions for their blameworthy acts. Members of the out-group will get the opposite treatment, their blameworthy acts being more likely to be attributed to their dispositions, their praiseworthy acts to the situation.

Social psychologists have coined the phrase “ultimate attribution error” to describe this very patterning. Taylor and Jaggi (1974) first investigated intergroup attribution in southern India, against the backdrop of Hindu-Muslim conflict. They asked Hindu subjects to imagine themselves in various situations with either a Hindu or Muslim interlocutor: a shopkeeper is generous or cheats you out of
Among Western participants only, the average asymmetry for negative events was reverse asymmetry for positive events was. In particular, attributions for negative events yielded the expected actor–observer asymmetry, \( d = 0.243 [0.135, 0.350] \), whereas positive events showed the reverse pattern, \( d = 0.149 [0.280, 0.019] \).
your change; you suffer a minor injury and someone helps or ignores you; you’re caught in the rain and someone offers or refuses you shelter; a teacher scolds or praises you. Subjects were then asked to explain the interlocutor’s action. They could choose from among explanations which involved internal attributions —like “Hindu shopkeepers are generous people,” and “Muslims tend to be rude and unreasonable”— and external attributions—like “he was in a hurry” and “he was constrained by social rules.” In all scenarios, Hindus were more likely to give internal attributions when they explained the socially desirable behavior of another Hindu agent; and in two of the four scenarios they were more likely to give internal attributions for undesirable actions when the actor was Muslim.

The study was replicated in Malaysia by Hewstone and Ward (1985) with Malay and Chinese subjects. They found that “Malays attributed a positive act by a Malay actor more to internal factors than a similar act by a Chinese actor, and they attributed a negative act by an in-group member less to internal factors than a similar act by an out-group actor” (p. 314). Similar attributional patterns have been observed with African American, Hispanic, and White, fifth and sixth graders in the Southwest of the United States (Stephan, 1977). Interestingly, attributions for Arab and Israeli actions during the Arab-Israeli war actually showed the same patterning for national actions, successes, and failures (Wolfsfeld and Rosenberg, 1977).

Given the well-established tendency of in-groups to regard their own members as, on average, morally better than out-group members (see Levine and Campbell 1972; Leach et al. 2007; Ellemers et al. 2008), intergroup attributional judgments pattern just as the mismatch theory predicts.
CHAPTER 2. THE MISMATCH THEORY

2.2 Modal explanations of attributional judgments

At a conceptual level, the modal and mismatch theories are strikingly different, but as we shall see, they generate many similar empirical predictions. Can we find an empirical difference between the two theories which goes beyond intuitions about cases like the ones presented earlier in this chapter?\(^4\)

Whereas the mismatch theory explains the drive toward situational or dispositional attributions in terms of attributers’ moral attitudes, the modal theory attempts to do so in terms of beliefs about how agents would behave were their circumstances somewhat altered. So in cases where attributional judgments vary independently of any plausible psychological difference in the causes of an agent’s behavior, the modal theorist must locate some other difference to explain the variation in her modal judgments.

To see the challenge, consider two people, A and B, observing one of Milgram’s subjects deliver painful shocks to the learner, and suppose that A but not B believes that the subject is a sadist. I submit, and the mismatch theory predicts, that A will judge the obedient behavior to be more person-caused, whereas B will judge it to

\(^4\)Imagine social scientists and marketers discover that the brand of stockings one prefers can depend on where the department store places it on the shelf (Chandon et al., 2009; Nisbett and Wilson, 1977). It would then be natural to say that one’s choice of stockings is sometimes situationally controlled. The modal theory can account for this judgment—the shelves could easily have been arranged otherwise, and had they been, one would have chosen differently—but since one’s choice was also evaluatively inconsequential, the mismatch theory has nothing to say about it.

Thus it may seem that cases of evaluatively inconsequential actions distinguish the two theories, and in such cases I grant that the modal theory provides a better account. However, as I hope these pages will show, the modal theory is inadequate when we consider morally valenced actions. For this reason I think that “the” person/situation distinction serves quite different purposes in moral and nonmoral contexts. Our concern here is solely with morally valenced cases. In §2.4, I will suggest that people draw an important distinction between dispositions that are grounded in an agent’s essence and those that are grounded in her more superficial features. The essential dispositions are especially salient when addressing certain moral questions, and it is our judgments about actions thought to arise from these dispositions that the mismatch theory is concerned to explain. For all I say here, some version of the modal theory may provide the correct account for how we think about superficial dispositions. Notably, however, the question of whether an action is self-disclosing is utterly out of place in morally neutral contexts, and the vast majority of studies in the attribution literature consider morally valenced actions (see Malle, 2006).
be more situation-caused. What explains their disagreement? The mismatch theory answers: B, but not A, believes the subject to be morally good, so only B expects the subject to disobey the experimenter. Thus, when A and B observe the subject’s obedience only B’s expectation that the moral status of an agent and her actions will match up is frustrated. A’s corresponding expectation is met.

The modal theorist’s task is less straightforward: she must explain why A’s and B’s beliefs about the relevant modal features of the subject’s psychology differ in virtue of their differing beliefs about whether she is sadistic. But such an explanation is not too hard to find. It is common knowledge that even quite large variations in the Milgram scenario will not deter a genuine sadist from torturing the learner. That a sadist’s cruelty is modally robust is just a bit of folk psychology. Hence, by drawing on folk psychology, the modal theorist actually can account for the judgment that the behavior of really sadistic subjects is person-caused.

Since one’s moral attitude towards obedience to authority will help to determine the extent to which one perceives the participants’ actions as morally good or bad, the mismatch theory predicts that such attitudes will modulate observers’ attributional judgments. This prediction may seem to fare better in distinguishing the modal and mismatch theories, for there is at least no obvious reason why the attributer’s moral attitudes towards obedience to authority should have a bearing on whether she expects an obedient behavior to be modally robust or fragile. Yet the modal theorist may be able to make sense of this variation as well. First, notice that evaluative attitudes are transparent. So when Bill, the pliant admirer of obedience, sees one of Milgram’s subjects obey the experimenter’s order to painfully shock the learner, what he sees is just someone doing what they ought to do because they, like him, know it’s what they ought to do—it’s what they’ve been ordered to do. Now Bill, like the rest of us, expects behavior to be more modally robust the more confidently it is believed to be what one ought to do. That’s just the platitude that
people are more apt to waver when they doubt that they’re doing as they ought. So compared to those of us who think that the subject ought to disobey, dutiful Bill will expect her to obey across a wider region of logical space. So again it seems that by drawing on a bit of folk psychology the modal theory too can make room for the attributer’s moral evaluation of the target behavior.

Another point of apparent distinction between these two accounts is the mismatch theory’s asymmetric treatment of good and bad agents. As we have seen, what it takes for a good agent to be situationally controlled is different from what it takes for a bad agent to be situationally controlled. This asymmetry does not obviously play any role in the modal theory. But it may receive much the same treatment as the role of the attributer’s evaluative attitudes received above. A good agent’s bad actions, not her good ones, are judged to be situationally controlled because we expect good deeds to reliably flow from a good agent; her bad deeds are indulgences, flukes, aberrations. Mutatis mutandis for a bad agent’s occasional good deeds. So here too the modal and mismatch theories appear to agree. It is proving difficult to see any daylight between them.

What can the modal theorist say about intergroup attribution biases, self-serving bias, and the fundamental attribution error? The biases can certainly be described in the terms of the modal theory, but what explains the patterning of attributions? As above, the modal theorist’s only plausible explanation draws on folk psychology: it is because we believe that we and ours are morally better than average that we expect ourselves to act morally over comparably wider regions of logical space and immorally over comparably narrower regions of logical space. Similarly, only by invoking folk psychology can the modal theory plausibly explain why moralists may be more sensitive to attributional biases, or why in-group/out-group attributions pattern as they do.

Is this a case of the moral tail wagging the psychological dog? The mismatch
and modal theories purport to explain the same phenomena and agree in many cases about which actions are situationally caused. An intriguing explanation for this consilience is that we rely on our moral attitudes for evidence when reasoning about how agents will behave in various counterfactual circumstances (see Figure 2.2). If the moral is in this sense evidentially prior to the modal, we should expect the two theories to be consilient.

2.3 Two new studies supporting the mismatch theory

We have seen that the overall patterning of the attribution literature and the results of intergroup attribution studies tend to support the mismatch theory. But we still do not know if the person/situation distinction fundamentally tracks how agents will act in various counterfactual circumstances, or if it fundamentally tracks attributers’ moral attitudes. And we still do not know if the connection between self-disclosure and the person/situation distinction is indirect, as it appears on the modal view (see Figure 2.1 on page 33), or if these two notions are really just different windows looking onto a single phenomenon, as the mismatch theory holds. The following two sections describe studies that I have conducted to help shed some light on these questions.

2.3.1 Study 1: Mark’s move

The mismatch theory holds that people’s moral attitudes play a central role in determining whether they explain an action in terms of situational or internal factors. Study 1 tests this hypothesis by looking for correlations between attributions and moral attitudes towards homosexuality.
CHAPTER 2. THE MISMATCH THEORY

Method

170 North American participants were recruited on Amazon Mechanical Turk (\(M_{\text{age}} = 27\) years, \(SD = 10.9\); 40% female; HIT approval rate \(\geq 95\%)\).

All participants read the following vignette:

Mark believes that homosexuality is morally wrong. In fact, Mark now leads a seminar in which he coaches homosexuals about techniques they can use to resist their attraction to people of the same-sex. However, Mark himself is attracted to other men. He openly acknowledges this to other people and discusses it as part of his own personal struggle.

At the end of one of his week-long seminars Mark visits a local bar. So by the time he notices that the handsome man sitting across the room has been smiling at him, Mark is feeling tired and has had a few drinks.

Mark walks up to the man and begins to come on to him.

Following Newman et al. (2013), manipulation checks asked “What does Mark believe?” (Homosexuality is wrong/Homosexuality is permissible) and “What

---

5 In a pilot study to determine appropriate exclusion criteria, I asked a small group of undergraduates to read and respond to the survey “briskly but thoroughly.” On average, they required 2:30 to do so. Nearly half of the mTurk sample completed the survey in under 2:15, often showing strong satisficing response patterns. For three of the five questions, selecting strongly agree indicated a negative attitude to homosexuality, while strongly agreeing with the remaining two indicated a positive attitude. Two items were therefore reverse coded: for these items, selecting strongly agree adds only one point to a participant’s score, rather than the usual five. One way of generating a score of 17 is therefore to select strongly agree for all five items. This unlikely response pattern was recorded over thirty times in Study 1. I conclude that these participants were probably just clicking through the survey as fast as possible. This is further supported by the fact that the ATLG is highly internally consistent (Cronbach’s \(\alpha \approx 0.9\)), and nearly all participants who scored 17 completed the survey in under 2:30. These participants were therefore excluded from the analysis. Appendix 4.3 presents analyses using alternative exclusion criteria.

6 The first half of this vignette is taken from Newman, Bloom, and Knobe (2013). Newman et al. investigated what drives people to locate a particular mental state (e.g., homosexual feelings or the belief that homosexuality is immoral) in an agent’s true self as opposed to her superficial self. They found that while liberals are most likely to respond that Mark’s homosexual feelings are his true self, conservatives are most likely to respond that the belief that homosexuality is wrong is his true self. The account which emerges is roughly that, of several conflicting mental states, an attributer will tend to describe the state which is morally best according to the attributer as most belonging to the agent’s true self.
does Mark experience?” (Attraction toward the same sex/Negative feelings about people who are attracted to the same sex). An additional check asked “Suppose that Mark went home with the man and had sex with him. How would he feel the next day?” (Pleased at his success/Disappointed with himself/Not fazed either way). 40 participants failed at least one check and were excluded from the analysis.

After reading the vignette, participants were asked to “indicate how important each of the following is to explaining why Mark came on to the man.” Two of the items gave non-agential explanations: (a) Mark was worn out from leading the seminar, and (b) The man was inviting Mark to flirt with him; and two gave internal explanations: (c) Mark felt attracted to the man, and (d) Mark wanted to sleep with the man. Participants rated each item using continuous sliders with extremes anchored at not at all important and extremely important. Since we are interested in the relative degree to which a subject favors or disfavors each explanation, I calculated absolute deviations for each score. Absolute deviations represent how individual importance ratings for each item deviate from the participant’s average across the four items. Items were counterbalanced for order.

Participants then completed the short form of the male version of the Attitudes Towards Lesbians and Gay Men (ATLG) scale (Herek, 1998). Participants rated items using 5-point Likert scales anchored at strongly disagree and strongly agree (Cronbach’s \( \alpha = 0.89 \); \( M_{\text{score}} = 11 \); \( SD = 6.5 \)). The items on the ATLG are: I think male homosexuals are disgusting; Male homosexuality is a perversion; Male homosexuality is a natural expression of sexuality in men; Sex between two men is just plain wrong; Male homosexuality is merely a different kind of lifestyle that should not be condemned. Items were counterbalanced for order.

---

\(^7\)Chapter 2.4 discusses substantive criteria for sorting explanations into the categories of internal and situational. However, even in lieu of a satisfactory account of the distinction, the analysis reported below indicates that the classifications assumed above have a solid basis in how people actually explain Mark’s action.
On a subsequent page of the survey, participants also responded to a self-disclosure prompt using a 9-point Likert scale anchored at strongly disagree and strongly agree: “Some of our actions reflect who we are deep down; they show our true selves. Please indicate how much you agree or disagree with the following statement: By flirting with the man, Mark showed who he most truly is deep down.”

Finally, participants identified themselves along a continuous scale ranging from extremely liberal to extremely conservative.

Results

The main result of this study is that ATLG scores are meaningfully correlated with locating greater importance in situational factors when explaining Mark’s action (see Table 2.1 on page 49). In other words, the more negative one’s attitude to male homosexuality, the more importance one will tend to assign to Mark’s being wiped out from teaching, and the less importance one will tend to assign to his erotic feelings when explaining his action.

Importance ratings correlated with political identification in the expected way for three of the items, though on the whole less strongly (see Table 2.1 on page 49). This is to be expected as political identification and ATLG scores are strongly but imperfectly correlated ($r = .63; p < .001$).

A principal component analysis on responses to items (a), (c), and (d), extracted a single component that models 62.4% of the variance in responses (Eigenvalue = 1.87; KMO Sampling Adequacy = 0.56—when all four items are included, KMO < .5; Bartlett’s Test of Sphericity < .001). Loadings for this component matched theoretical expectations, and component scores correlated in the predicted way with ATLG scores ($r = .59 [0.41 - 0.72]; p < .001$) and political identification ($r = .28 [0.05 - 0.48]; p = .02$).
The mismatch theory’s predictions about self-disclosure judgments were also confirmed, corroborating Newman et al.’s (2013) findings. For ATLG scores and self-disclosure, $r = -0.59 [-0.73, -0.42]$ ($p < .001$); and for conservativeness and self-disclosure, $r = -0.37 [-0.55, -0.15]$ ($p = .002$).

**Discussion**

When asked to predict the results, people rarely have a clear sense that homophobic participants will be more inclined to think, for example, that Mark’s being wasted from teaching is important to explaining his action. (Homophobic participants hardly seem likely to empathize, “Well, if I’d been that tired from teaching I’d have come on to him too.”) But the more negative one’s attitude towards homosexuality, the more one will perceive a conflict between the moral valence of Mark and the moral valence of his action, so this seemingly odd result is just as the mismatch theory predicts.

The significant correlation between ATLG scores and participants’ inclination toward situational explanations suggests that the mismatch theory does not merely describe how attributions will pattern into the broad categories of internal and situational; it also describes, to a fair approximation, the degree to which they will so pattern. These results therefore appear to provide further support for the view.

It may be objected that homophobic participants give more situational attributions for Mark’s action not because of their moral attitudes towards homosexuality but rather because they are inclined to think that had Mark been a little less tired, or had he not topped off that last beer, or ... he would have resisted his feelings. Whereas conservatives and people with high ATLG scores are inclined to think that Mark’s homosexuality is fundamentally something he has chosen, liberals and people with low ATLG scores are more inclined to think that he was born gay and there’s really not much use fighting it (see, e.g., Haider-Markel and Joslyn
Table 2.1: Correlations of attitudes to gay men (ATLG scores) and conservativeness with importance ratings for items (a) - (d) and component scores for items included in factor analysis (extraction method: principal component analysis). ‘PCA loading’ indicates the component loadings (i.e., the correlations between each item and the principal component).

<table>
<thead>
<tr>
<th>Item</th>
<th>$r$</th>
<th>$p$</th>
<th>95% CI</th>
<th>PCA loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>ATLG</td>
<td>0.54</td>
<td>&lt;.001</td>
<td>0.39, 0.74</td>
</tr>
<tr>
<td></td>
<td>CONS.</td>
<td>0.19</td>
<td>.10</td>
<td>-0.04, 0.41</td>
</tr>
<tr>
<td>(b)</td>
<td>ATLG</td>
<td>0.03</td>
<td>.80</td>
<td>-0.20, 0.26</td>
</tr>
<tr>
<td></td>
<td>CONS.</td>
<td>0.15</td>
<td>.22</td>
<td>-0.09, 0.64</td>
</tr>
<tr>
<td>(c)</td>
<td>ATLG</td>
<td>-0.33</td>
<td>.005</td>
<td>-0.53, -0.11</td>
</tr>
<tr>
<td></td>
<td>CONS.</td>
<td>-0.23</td>
<td>.06</td>
<td>-0.44, 0.00</td>
</tr>
<tr>
<td>(d)</td>
<td>ATLG</td>
<td>-0.50</td>
<td>&lt;.001</td>
<td>-0.66, -0.30</td>
</tr>
<tr>
<td></td>
<td>CONS.</td>
<td>-0.25</td>
<td>.04</td>
<td>-0.45, -0.01</td>
</tr>
</tbody>
</table>

Component

<table>
<thead>
<tr>
<th>Item</th>
<th>$r$</th>
<th>$p$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATLG</td>
<td>0.59</td>
<td>&lt;.001</td>
<td>0.41, 0.72</td>
</tr>
<tr>
<td>CONS.</td>
<td>0.28</td>
<td>.02</td>
<td>0.05, 0.48</td>
</tr>
</tbody>
</table>

Figure 2.1: Plot of ATLG scores vs. overall attribution inclination ($r = .53$, $p < .001$).
2008; Lewis 2009; Suhay and Jayaratne 2012; Jayaratne et al. 2006). The objector concludes: these results (and perhaps Newman et al.’s) can be explained on the plausible hypothesis that people are inclined to explain actions that seem to arise from innate dispositions in terms of agential rather than situational factors.

2.3.2 Study 2: Kate’s pregnancy

In light of the evidence in favor of the mismatch theory, it seems plausible that the best explanation for the patterning of judgments in Mark’s case is that conservatives and participants with high ATLG scores are more inclined than liberals and participants with low ATLG scores to disapprove of homosexuality. Still, the fact that people who think male homosexuality is morally wrong may be more likely, relative to base rates, to think that it is “a choice” does muddy the water, so it would be good if we had a case which avoided this potential confound. The following study aims to address this worry.

Method

160 North American participants were recruited using Amazon’s Mechanical Turk service ($M_{\text{age}} = 33, SD = 10; 41\%$ female; HIT approval rate $\geq 95\%$).

All participants read the following vignette:

Kate is a senior at college. Just like many of her friends, after graduation Kate plans to spend a year working for a charity organization before pursuing her dream of going to medical school. Kate has recently discovered that she is pregnant. And what’s worse, her boyfriend broke up

---

822 participants failed at least one of two comprehension checks and so were excluded from the analysis. In a pilot study, a small group of undergraduates were asked to read and respond to the survey “briskly but thoroughly.” On average, they required 2:15 to do so. 61 participants in the online sample who passed the manipulation checks still managed to complete the survey in under two minutes, again often showing strong satisficing response patterns. These participants were therefore excluded from the analysis. The complete data set is available online <bit.ly/mismatch_kate> and analyses of the complete data set are presented in Appendix 4.3 on page 108.
with her just a week earlier. When she tells her parents, their reaction is clear: although they will support her no matter what she decides, they both think that she should abort the pregnancy. Kate’s college health clinic offers the procedure at no charge. After thinking it over, Kate decides to have an abortion.

Comprehension checks asked: “Suppose that Kate decided not to get an abortion, how would her parents probably react?” (They would disown her / They would be concerned, but would ultimately stand by her / They would be pleased), and “What is Kate planning to do after she graduates?” (Study to become a registered nurse / Work for a charity / Take a gap year before looking for a job).

As in Study 1, after reading the vignette participants used sliders with extremes anchored at not at all important and extremely important to indicate how important they thought various features of the case are to explaining the agent’s action. Two items cited internal causes: (a) Kate believes it is not the right time for her to start a family, and (b) Kate wants to begin her career without the burden of a young child; and two items cited situational causes: (c) Kate can easily get an abortion at her College’s health clinic, and (d) Kate’s parents pressured her to have an abortion. I converted individual importance ratings for each item into absolute deviations.

Participants then indicated the extent to which they agreed or disagreed with each of the following: Having an abortion is a disgusting thing to do; Abortion is just plain wrong; Access to abortions should be free and easy; Doctors who perform abortions are evil; Abortion is only ever justified when a pregnancy seriously threatens the mother’s life. Responses to this scale were highly internally consistent (Cronbach’s $\alpha = 0.89$). Since participants used a 7-point Likert scale, the maximum possible score was 35 and the minimum was 5. The mean was 16.6 ($SD = 8.1$). All items were counterbalanced for order.
On a subsequent page of the survey, participants indicated the extent to which they agreed or disagreed with the statement, “Kate’s decision to get an abortion reflects what she wants deep down.” (I arrived at this phrasing after pilot studies reminded me that statements like “When Kate gets an abortion, we see the real Kate” have a distinctly punitive ring in this context.) Finally, participants identified themselves along a scale from extremely liberal to extremely conservative.

Results

The main result of this study is just as the mismatch theory predicts: the more negative one’s attitude toward abortion, the more inclined one will be to locate the explanation of Kate’s action in situational factors; the more positive one’s attitude toward abortion, the more one will be inclined to locate the explanation of Kate’s action in Kate herself (see Table 2.2 on the next page). Importance ratings for all items also correlated with political identification in the expected ways.

Principal component analysis extracted a single component that models 73.0% of the variance in the responses to items (a), (b), and (d) (Eigenvalue = 2.18; KMO = 0.70—with all four items included, KMO = 0.21; Bartlett’s Test < .001). Item loadings for this component again closely matched theoretical expectations, component scores correlated in the expected way with attitudes to abortion ($r = .65$ [0.49 - 0.76]; $p < .001$) and political identification ($r = .38$ [0.17 - 0.55]; $p < .001$), suggesting that negative attitudes toward abortion are significantly correlated with placing relatively more weight in situational explanations of Kate’s action.

Responses to the self-disclosure prompt also patterned in the expected way: the more negative one’s view of abortion, the more one will tend to think Kate’s decision does not reflect what she wants deep down. For attitudes to abortion and self-disclosure, $r = -.45$ [-0.61, -0.25] ($p < .001$); for conservativeness and self-disclosure, $r = -.21$ [-0.42, 0.01] ($p = .06$).
Table 2.2: Correlations for attitudes to abortion (ATA) and conservativeness (CONS.) with importance ratings for items (a) - (d) and component scores from principal component analysis. The statistics described in this table correspond to those described in Table 2.1 on page 49.

<table>
<thead>
<tr>
<th>Item</th>
<th>r</th>
<th>p</th>
<th>95% CI</th>
<th>PCA loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>ATA</td>
<td>-0.56</td>
<td>&lt;.001</td>
<td>-0.70, -0.39</td>
</tr>
<tr>
<td></td>
<td>CONS.</td>
<td>-0.41</td>
<td>&lt;.001</td>
<td>-0.58, -0.20</td>
</tr>
<tr>
<td>(b)</td>
<td>ATA</td>
<td>-0.52</td>
<td>&lt;.001</td>
<td>-0.66, -0.33</td>
</tr>
<tr>
<td></td>
<td>CONS.</td>
<td>-0.21</td>
<td>.07</td>
<td>-0.41, 0.02</td>
</tr>
<tr>
<td>(c)</td>
<td>ATA</td>
<td>0.25</td>
<td>.03</td>
<td>0.03, 0.45</td>
</tr>
<tr>
<td></td>
<td>CONS.</td>
<td>0.13</td>
<td>.27</td>
<td>-0.10, 0.34</td>
</tr>
<tr>
<td>(d)</td>
<td>ATA</td>
<td>0.57</td>
<td>&lt;.001</td>
<td>0.39, 0.70</td>
</tr>
<tr>
<td></td>
<td>CONS.</td>
<td>0.34</td>
<td>.002</td>
<td>0.13, 0.53</td>
</tr>
<tr>
<td>Component</td>
<td>ATA</td>
<td>0.65</td>
<td>&lt;.001</td>
<td>0.49, 0.76</td>
</tr>
<tr>
<td></td>
<td>CONS.</td>
<td>0.38</td>
<td>&lt;.001</td>
<td>0.17, 0.55</td>
</tr>
</tbody>
</table>

Figure 2.2: Study 2. Plot of scores on the attitudes to abortion scale vs. overall inclination to situational attributions (r = .62, p < .001).
CHAPTER 2. THE MISMATCH THEORY

Discussion

As with Study 1, these results can seem surprising. It isn’t at all obvious what effect one’s moral attitude toward abortion will have on how much explanatory weight ones places on, for example, Kate’s belief that it’s not the right time for her to start a family. Without the insight provided by the mismatch theory, these results might appear rather inexplicable.

Moreover, these results cast doubt on the objection to Study 1. In lieu of some reason to think otherwise, both studies should be explained in the same manner; but the objector’s hypothesis—that people’s modal beliefs covary with the relevant moral attitudes—though indeed plausible in Mark’s case, is rather implausible here. It seems, then, that it is participants’ moral attitudes towards homosexuality, and not their beliefs about whether homosexuality is a choice, that explains the patterning of attributions for Mark’s action.

2.3.3 Ego-syntonic/Ego-dystonic

The foregoing experiments also point to the problem with Sabini, Siepmann and Stein’s important analysis of the person/situation distinction (2001; also see Sabini and Silver, 1987). They claim, correctly, that the distinction has nothing to do with the boundary of the epidermis or covariation with either the person or the situation. On their view, the distinction between person-caused and situation-caused actions is “between causes of behavior people affirm as part of themselves and causes they reject,” that is, “between ego-syntonic and ego-dystonic causes,” respectively (p. 11).

Naturally, I think this analysis is importantly on the right track, but as the results of the present studies suggest, it is not psychologically adequate and it cannot capture our intuitions. For both liberals and conservatives agree that Mark’s action is not caused by something he affirms as part of himself—he “does not want to be
CHAPTER 2. THE MISMATCH THEORY

the sort of person” (ibid.) who hits on men. But not everyone agrees that his action is caused by the situation, as Sabini et al.’s account predicts they should.

Other cases make the point even more forcefully. Consider Jonas:

Above all else, Jonas values the purity of the Aryan race, and he relishes every part of his hard-won job in the S.S. Occasionally, however, Jonas finds that he has deep and overwhelming feelings of sympathy for some of the people he is required to kill. But these occasions usually don’t last long: most of the time he comes around and kills more people.

One day, Jonas is ordered to execute a large Jewish family, including several small children. He firmly believes that killing this family is the morally right thing to do, but when it comes time to act he is overwhelmed by his powerful sympathetic feelings. He cannot force himself to do what he believes is morally right. He regretfully lets them escape to safety.

The immediate reaction to Jonas’ merciful action is that it is person-caused. But it is ego-dystonic: it is inconsistent with his second-order desires and his most deeply held beliefs; it is external to who he would be; he does not affirm its cause as a part of himself, and all the rest. So Sabini et al.’s view incorrectly classifies it as situation-caused.

The point has a direct parallel in the philosophical literature on responsibility and self-disclosure: it is the attributional version of what Arpaly and Schroeder (1999) label inverse akrasia. A person acts akratically when she acts against her best judgment: she believes that she has most reason to do one thing, but does something

---

9When asked “Does the cause of Jonas’ merciful action lie more within Jonas himself, or more within the situation Jonas is in?” the average response on a 9-point scale (1 = more within Jonas himself; 9 = more within the situation) was 3.1 [2.38, 3.84] (SD = 2.25). Responses to this scale were strongly correlated with self-disclosure responses (r = .63, p < .001). (Age_{av} = 29.2 (SD = 8.6), 48% female).
else all the same. In many cases, akratic agents act badly (the kleptomaniac, the unwilling addict, etc.), but as Arpaly and Schroeder point out, weak-willed actions can also be morally good. Consider their favorite example, Huckleberry Finn. Huck believes he has most reason to turn the escaped slave, Jim, over to the authorities, and since he wants to be a good boy, he desires to turn Jim in and wants for this desire to effectively guide his action. But he simply cannot bring himself to act as he believes he should—his sympathetic feelings for Jim are too strong to resist.

No matter whether we understand reflective endorsement in terms of Huck’s second-order desires or in terms of his values, Huck does not reflectively endorse allowing Jim to escape—he wishes he could overcome the sympathetic feelings that make him incapable of acting in accord with his racist values. But far from threatening his responsibility, Arpaly and Schroeder take this to be “indicative of the fact that Huckleberry is, in an important sense, a good boy, a boy with his heart in the right place” (p. 165, my emphasis). We might add that Huck’s benevolent action does not seem to arise from the situation in which he finds himself; rather, its cause seems to lie deep within Huck himself, just as the mismatch theory predicts.

Since it is a straightforward application of philosophical ideas about agency and identification, Sabini et al.’s account suffers from a parallel problem to that highlighted by the case of Huck Finn. Consider what they say about “the addict’s” behavior (cf., Frankfurt 1971):

[It] is externally caused . . . not because his cravings are external to either his skin or his consciousness, or, heavens knows, his dispositions—but because he rejects them as part of himself. They are foreign to his sense of self, certainly not because he lacks a disposition to take whatever he is addicted to, but because his desire, his disposition, is itself external to who he would be (p. 12).
Sabini et al. are right that people understand situation-caused actions as “external not to the person, but to the person’s self,” but the self must be understood as the true self, and what belongs to a person’s true self is not simply a matter of what that person affirms as a part of herself—or indeed, of any other naturalistic feature of her psychology.

Sabini et al. tend to be a bit fast and loose in their expression. Initially, they say that “behavior is internally caused if and only if it follows from a person’s values and (correct) beliefs,” and then go on to elaborate this point in many ways. They say that situation-caused actions are “ego-dystonic,” “foreign to [the person’s] sense of self,” “external to who [the person] would be,” inconsistent “with what the person expressing the attitude really believes,” “consistent with the person’s regrettable dispositions,” and “contrary to second order desires.” These formulations are not equivalent, but none of them will do; for everyone judges that Mark’s action is foreign to his sense of self, external to who he would be, inconsistent with what he believes, consistent with dispositions he regrets, and contrary to his second-order desires. But for all that, it still seems (to low-prejudice people) that Mark’s hitting on the man arises not from the situation, but from within Mark himself.

So I agree that we tend to think of the “causes of behavior people affirm as part of themselves” as self-disclosing and dispositional. However, as Jonas’, Mark’s, and Huckleberry Finn’s cases suggest, this is not an invariable pattern. Even though Mark’s religiously motivated beliefs are ego-syntonic, liberals doubt that they are really a part of his true self. Correlatively, even though Jonas’ sympathetic feelings are ego-dystonic, they seem to lie within his true self.

It seems, then, that we will discount an agent’s own affirmations when we strongly disapprove of her ego-syntonic attitudes (as most of us do in Jonas’ and Huck Finn’s cases), or when an ego-dystonic attitude of which we approve conflicts with an ego-syntonic attitude of which we disapprove (as it seems to low-prejudice
people in Mark’s case). In other words, we believe that people can be mistaken in what they affirm as a part of themselves: a person’s true self is something she must discover. Sabini et al. treat as constitutive what is, at best, merely heuristic.

2.3.4 A Watsonian challenge

If the picture of the folk psychology of attribution developed here is correct, judgments about self-disclosure cannot serve as a non-moral basis for judgments about whether an agent is blameworthy, for self-disclosure judgments are themselves morally laden. The evidence is not conclusive, however, so perhaps there are sustainable alternative interpretations of the data that might leave folk attribution theory in better shape. Of the common philosophical accounts, Gary Watson’s insightful analysis of “deep attributability” seems particularly well-suited to explain the apparent patterning of people’s judgments:

The significant relation between behavior and the “real self” is not (just) causal but executive and expressive . . . [I]f what I do flows from my values and ends, there is a stronger sense in which my activities are inescapably my own: I am committed to them. As declarations of my adopted ends, they express what I’m about, my identity as an agent. They can be evaluated in distinctive ways . . . because they themselves are exercises of my evaluative capacities (p. 233).

To return to the case of Mark, a Watsonian might interpret the findings presented here as showing merely that liberals and conservatives disagree about whether Mark’s action flows from the exercise of his fundamental evaluative capacities—that is, from his values.

Suppose, as seems plausible, that we are inclined to believe that most morally good people share our values (Ross et al., 1977). Then the more one believes that
homosexuality is seriously morally wrong, the more one will tend to believe that
Mark is evaluatively committed to resisting his attraction to the man—given that
one believes Mark is a morally good person. It follows that the more one believes
that homosexuality is seriously morally wrong, the less one will tend to believe
that Mark’s hitting on another man is expressive of his values. On the Watsonian
analysis, this is just to say that one will tend to judge that Mark’s action is not self-
disclosing. Moreover, if we think that morally bad people do not share our values,
this approach might even explain why the pattern reverses when we consider
morally evil agents.

The evidence presented here may not, therefore, appear to raise any new diffi-
culties for the idea that self-disclosure is a normatively significant action-theoretic
relation. And given the plausible assumption that we project our own values onto
people whom we judge to be good, it might actually support a Watsonian account
of self-disclosure.

This is a good challenge. I certainly agree that people have evaluative commit-
ments and that some, but not all, of their actions will reflect those commitments.
So I agree that the Watsonian has characterized a genuine action-theoretic relation,
and moreover, that there is no reason to adopt an error theory about this relation.
It is coherent and sometimes instantiated. (It suggests, moreover, an extremely
interesting account of the person/situation distinction.) But for this hypothesis to
explain the empirical data, it must be true that liberals and conservatives tend to
disagree about what Mark values, and there is no evidence that they do. It seems,
rather, that both liberals and conservatives agree that Mark’s anti-gay mission is
among his fundamental evaluative commitments; they certainly agree that he be-
lieves homosexuality is morally wrong and that he would feel terribly guilty were
he to act on his erotic feelings (see note 5). But if they agree that hitting on a man
is not expressive of Mark’s values, then the Watsonian account does not explain
why self-disclosure judgments appear to vary systematically with people’s moral attitudes towards homosexuality.

Moreover, if people think of values as genuine beliefs about the good (as Watson does), then presumably it is just obvious to both liberals and conservatives that hitting on a man is not expressive of Mark’s values. Thus Mark’s action is clearly not self-disclosing in the Watsonian sense. But as we have seen, liberals actually tend to think that Mark’s hitting on the man is self-disclosing, so the Watsonian account fails to predict their judgments.

2.3.5 The upshot for self-disclosure

To my knowledge, social scientists have not examined judgments about self-disclosure. However, it seems enormously plausible that if these researchers had asked about whether actions really belong to agents or reflect their true selves, the pattern of results would have been largely unchanged—they would have found exactly the same biases and asymmetries. It is, for example, enormously plausible that the Hindu participants in Taylor and Jaggi’s (1974) study would have been more inclined to judge that a Muslim shopkeeper, as opposed to a Hindu shopkeeper, reveals her true self when she shortchanges them, that her action really belongs to her, and all the rest. And just as with attributions, the pattern would have reversed when they considered morally good actions: Hindus would be more inclined to view the kind actions of other Hindus as self-disclosing than they are the comparable actions of Muslims.

If I have succeeded in convincing you that the mismatch theory is a plausible account of the principles that drive us toward person or situation attributions, you should seriously entertain the hypothesis that it is equally good as applied to self-disclosure. It seems that judgments about whether an action is sourced in an agent’s values, or backed-up by second-order volitions, or responsive to judgment-sensitive
attitudes, are not what drive us to say that the ordinary thief’s behavior, unlike the kleptomaniac’s, really belongs to her or reflects her true self. It is, rather, that we have different moral attitudes toward ordinary thieves and kleptomaniacs. We can make this vivid by imagining a liberal philosopher and a conservative philosopher arguing about whether Mark reveals his true self when he attempts to chat up the man at the bar. Our liberal philosopher holds that Mark’s case is a counterexample to the conservative’s analysis of self-disclosure: “Mark has erotic thoughts about the man, but he does not desire that they lead him to action effectively. Indeed, he experiences his own action as a kind of weakness, so there is no doubt that he does not have the will that he wants! But there is equally no doubt that his action reveals his true self, so your account must be false.” Our conservative philosopher disagrees, since to her it seems clear that Mark’s action was not self-disclosing.

It is hard to believe that their disagreement can be captured by any of the common philosophical analyses of self-disclosure. Rather, it seems that while they both think that Mark is a good person, their attitudes towards homosexuality differ. Thus only the conservative’s expectation that the moral valence of Mark and his behavior will match up is frustrated; that is what leads our two philosophers to contradictory judgments about whether Mark’s action reflects his true self. I maintain that it is the correct model for all judgments about whether actions are self-disclosing.

2.4 Interpreting the mismatch theory

The aim of this section is twofold: first, to provide further support for the mismatch theory by locating it in an established approach to social cognition; and second, to sketch an account of the content of our beliefs about people’s moral valences. But since we seem to have a tolerably good pre-theoretic sense of what these beliefs
amount to, before getting into the swing of things I should pause to motivate the need for such an account.

2.4.1 The need for an account of moral valence

The mismatch theory assumes that we have beliefs about the moral valences of agents, but it says little about the contents of these beliefs. At first, this assumption may appear unproblematic. For we seem to have a tolerably good sense of what we mean when we say that S is a good (or bad) person: we mean that S has a good (or bad) character. *To be a good person is to have a good character*. Call this claim the *characterological view* of people’s moral valences. The characterological view seems plausible both metaphysically and as an account of what the folk believe.\(^{10}\) So why do I insist that we require an account of our judgments about people’s moral valences—and moreover, as we shall see, an account with empirically and metaphysically implausible presuppositions?

By way of motivating the need for this chapter, I submit the following *a priori* puzzle for the characterological view. It appears to be a minimal requirement of having a morally good (or bad) character that one be disposed to do good (or bad) things over a fairly wide range of possible circumstances. That is, having a moral character of a certain sort is conceptually linked to being disposed to behave in certain ways in a broad range of circumstances. But this does not seem to be how we think about being a good person deep down. We seem to think that a person’s true self can be regularly obscured by her behavior in a way that her character simply cannot. If, for example, she is not reliably disposed to behave bravely in circumstances which call for bravery, it follows that she is not a brave person. More generally, if she is not disposed to behave virtuously, it follows that she does not

---

\(^{10}\)It is contentious whether this view has false empirical presuppositions; see Doris, 2002; Harman, 1999, 2000.
have a good character. But we think she may, for all that, still be good deep down. Thus parents might say of their child: “She is a violent criminal, and we admit, she does not have a morally good character. But we know in our hearts that deep down she is a good person.” What they say may be false, but it does not sound (at least to my ear) incoherent.\footnote{Indeed, this appears to be a fairly common literary trope. Danila Bagrov (from Sergei Sergeyevich Bodrov’s \textit{Brother}), Sandor ‘the Hound’ Clegane (from \textit{Game of Thrones}), George Costanza (from \textit{Seinfeld}), Larry David, Falstaff, Humbert Humbert (from Nabokov’s \textit{Lolita}), Gareth Keenan/Dwight Schrute (and perhaps David Brent, from Ricky Gervais’ \textit{The Office}), Macbeth, Maria Braun (from Fassbinder’s \textit{The Marriage of Maria Braun}), Wikus van de Merwe (from Peter Jackson’s \textit{District 9}), Homer Simpson, Anakin Skywalker, Tony Soprano, Freddie Quell (played by Joaquin Phoenix in \textit{The Master}), and Walter White (from \textit{Breaking Bad}), among others, have all been suggested to me as examples of people who are good deep down, but do not have morally good characters. Many of these examples also cast doubt on the idea that we are natural Kantians, i.e., that our judgments of people’s moral valences aim to describe the moral quality of their wills alone.}

Thus when we assert that S is a morally good person we cannot mean that S has a morally good character, although we may imply this pragmatically. Indeed, one plausible function of adding the qualifier “deep down” is precisely to cancel this implicature. But if we do not think that being a good person is primarily a matter of being disposed to do good things, or even to think good thoughts, what is it a matter of? Psychological essentialism gives a very bare answer: it is a matter of having a morally good essence. The following two sections consider (impressionistically) what role beliefs about essences might play in folk attribution theory.

\subsection*{2.4.2 Psychological essentialism}

Psychological essentialism is the theory that the folk implicitly believe that “a wide range of entities have within them a substance-like [essence], which causally grounds their most important, stable and enduring properties” (Leslie, 2013). Preschoolers, for example, seem to think that it is a girl’s essence that disposes her to wear dresses (Gelman, 2003, p. 95), but presumably it is something more superficial that disposes her to wear this particular dress with polka dots.
It is worth emphasizing that psychological essentialism is a fundamentally different view to the metaphysical essentialism familiar to philosophers. The fact that being striped is a part of a tiger’s essence (in the sense of psychological essentialism) does not imply that albino tigers are impossible; and the fact that being four-legged is a part of a dog’s essence does not imply that having a leg amputated is necessarily fatal for a dog. Psychological essentialism, unlike metaphysical essentialism, does not hold that every member of a kind necessarily shares the essential properties of that kind: like the albino tiger’s stripes, some may never appear, and like the dog amputee’s four-leggedness, some may be lost. Psychological essentialism holds only that human beings implicitly believe that many things have a hidden essence that guides their development and causally undergirds many of their most important properties (especially their kind-typical properties), and that many of the observable differences between kinds are caused and explained by differences in their hidden essences. (For a detailed discussion of the relation of psychological and metaphysical essentialism, see Leslie, 2013.)

Psychological essentialism is not committed to the claim that the folk believe in universal essences—essences that all members of a kind share exactly. Rather, the essentials in question might all be individual essences. Within a kind, individual essences are, of course, importantly similar, but they can differ in ways that do not bear on kind membership. (This is suggested, among other things, by the fact that individuals often belong to multiple, distinct kinds.)

Evidence for psychological essentialism has been steadily accumulating for the past quarter-century and the theory is now amply supported (see Gelman, 2003 for a thorough review of the literature to that date; Keil, 1989; Sobel et al., 2007; Newman and Keil, 2008; Gelman and Ware, 2012; Leslie, 2013, and the references therein). Although psychologists don’t put it this way, we now know that dispositions play two distinct roles in folk psychology: first, there are essential dispositions which
guide an organism’s development, are important to kind-membership, and stable across time; and second, there are superficial dispositions which are unrelated to development, variable within kinds, and malleable across time.

This folk distinction suggests a natural and extremely interesting interpretation of folk attribution theory: perhaps when people say that an action is situationally caused or that it fails to be self-disclosing they are really giving voice to their implicit belief that it was not appropriately related to the agent’s essence. On this interpretation, folk attribution theory is but another manifestation of essentialist thinking. There are several ways to spell out this intriguing possibility.

An essence-causal account

Let’s begin with the simple view that the only inputs to folk attribution theory are implicit beliefs about whether an action is caused by an agent’s essence. On this view, it is not obvious what role the attributer’s moral attitudes might play in generating attributional and self-disclosure judgments. But as we have seen, the vast literature on the actor-observer asymmetry, intergroup attributions, and the original empirical studies presented here, all suggest that moral attitudes play an important role in the folk theory of attribution.

One way of squaring this evidence with the causal interpretation is to treat the mismatch theory as describing something like a heuristic. We cannot directly perceive whether actions are caused by essences or by something more superficial, so we must rely on indirect evidence. We believe that good essences generally cause good actions, so when a person we believe to have a good essence does a bad thing, we have evidence that her action was caused by something other than her essence. Moral attitudes then feed into folk attribution theory as one source of evidence that an action might not have been caused by an agent’s essence.
Reflection on cases in which well-meaning agents who, in the pursuit of valuable ends, inadvertently bring about disastrous effects, suggest that this account is unlikely to be the whole story. (Consider, for example, a person who burns down the family home while stringing up holiday lights. Their actions—the bodily movements that led to the house fire—may be caused by their essential dispositions, but they do not seem self-disclosing. Actions which turn out badly due to luck rarely seem self-disclosing and often get situationally explained, independently of anything which we might plausibly take to bear on their causal etiology.

An essence-congruity account

A second and possibly complementary way to square the importance of moral attitudes with the idea that the folk theory of attribution is a manifestation of psychological essentialism takes the important relation between actions and essences to be congruity. On this view, the inputs to folk attribution theory concern not whether actions are caused by essences, but whether they are congruous with them. If we believe that people’s essences are morally valenced, then for an action to be congruous with an essence is, in part, for it to share the moral valence of the essence. Thus moral attitudes are not mere heuristics; they are among the fundamental inputs to folk attribution theory.

Do we implicitly believe that people’s essences have a moral quality of their own? Plausibly, yes. Apart from the obvious fact that many languages provide ways to express beliefs about whether people are (as we say it) good or bad at heart or deep down, recent empirical work also points in this direction. In one study, Strohminger and Nichols (2014) asked subjects to rate how much a person would change if she were to take a pill that selectively and permanently altered one part of her mind (p. 162). Morally significant traits (e.g., psychopathy, pedophilia, being a jerk) were consistently judged to cause the greatest changes. On average, moral
traits had a considerably larger impact on identity judgments than did personality traits, memories, preferences, and perceptual abilities. In other studies, moral traits were judged more likely to be transmitted to a new body along with a person’s soul, more likely to survive reincarnation, and more likely to persist over a forty year stretch of a person’s life, relative to base rates in each case. This evidence fits the profile of essentialist thinking more broadly: moral traits are important to identity, stable, and enduring.

The hypothesis that we implicitly believe in morally valenced essences generates an extremely surprising prediction: people should be averse to receiving body transplants from morally bad donors because they believe that contact with the morally bad donor’s essence will cause them to behave immorally. This prediction follows from two central features of essences: they are suffused (physically) throughout our insides and they causally undergird our important traits (Sobel et al., 2007).

Meyer, Leslie, and Gelman (2013) confirm this prediction, providing striking evidence for a belief in morally valenced essences. Meyer et al. report that both American and Indian subjects are averse to receiving body transplants from morally bad donors, and this is not because subjects believe such donors are more likely to be unhealthy, or because they believe knowing that they’ve received such a transplant will make them behave less morally. Rather, people seem to fear that receiving a body transplant from a bad donor will infect them with the donor’s morally bad essence (also see Itajkura et al., 2011, for studies with British and Japanese subjects). This, of course, presumes that the donor’s essence is morally bad, and a fortiori, that it has a moral quality of its own. This appears to be an example of the more general phenomenon of moral contagion: people are “extremely uncomfortable with the idea of coming into contact with personal items that belonged to individuals deemed morally negative (e.g., Hitler’s sweater)” (Meyer et al., 2013, p. 673). We appear to believe implicitly that these items have been contaminated with their
owner’s essence, and consequently, that contact with them can cause us to become immoral (Eskine et al., 2013; Gelman et al., 2014).

These studies add to the case mounted in §2.4.1 for the claim that judgments about an agent’s moral valence are probably not metaphysically light-weight judgments about her character. For, the hypothesis that we believe people’s moral valences are grounded in their essences helps to explain, first, why moral traits are thought to be transmissible via contact with people’s insides (e.g., via organ transplants); and second, why moral traits are thought to be especially stable, enduring, and important to people’s identities. It is unclear how, if at all, these findings can be explained on a characterological account of our judgments about people’s moral valences.12

If, as these findings suggest, we believe that people’s essences are morally valenced, the mismatch theory can be nearly taken at face value: we are inclined to judge that an action is situationally caused and not self-disclosing to the extent that the perceived moral valence of the action differs from the perceived moral valence of the agent’s essence. This suggests that the contents of self-disclosure judgments concern whether actions are appropriately related to essences, and that judgments about agents’ moral valences aim to describe the moral quality of their essences.

When we detect a mismatch in the moral valence of an action and the actor’s essence, we become inclined to look for situational explanations of the action. But what exactly is it that we’re looking for? The present interpretation suggests that we look for properties that play some role in explaining the action, but do not involve or otherwise implicate the person’s essence. These properties are likely to be unstable, ephemeral, and unimportant to identity. An explanation is situational, on this view, to the extent that it is confined to such nonessential properties.

12Adam Lerner suggested to me that the psychological essentialist account of people’s moral valences also helps to explain our reluctance to forgive: since essences are thought to be stable and enduring, if we come to believe that S is a bad person (for example, by being wronged by S), we will find it difficult to revise that belief, which will in turn make it difficult for us to forgive S.
This account generates empirically testable hypotheses about how people will explain behavior, and further arguments may be adduced in its favor. But I shall not pursue those tasks here.

### 2.5 Conclusion

What are the consequences of accepting the mismatch theory? If you think that attributional judgments concern whether actions are appropriately related to essences, you should seriously consider an error theory about the person/situation distinction and self-disclosure. That is, you should think that the language of folk attribution theory attempts to describe the world, but fails to do so since there is nothing in human beings to play anything like the role of an essence (Leslie, 2013). The person/situation distinction and self-disclosure judgments presuppose the same empirical falsehoods that lead many people to believe that receiving a heart transplant from a violent murderer will cause them to act more like violent murderer. It seems to follow, then, that an uncorrected folk notion of self-disclosure should play no role in normative accounts of moral responsibility.

It may be objected that since philosophers mean to provide normative analyses of how we ought to think about self-disclosure—and nothing about how we ought to think follows directly from facts about how we do think—philosophical analyses of self-disclosure are insulated from the kinds of empirical concerns raised here. This response is undoubtedly onto something important, but I do not find it wholly satisfying for two reasons. First, we use our intuitive judgments about which actions

---

13If you think that moral thought and talk serve primarily to express noncognitive attitudes—for example, approval and disapproval—then you should consider an expressive analysis of the person/situation distinction. If moral judgment does not even purport to describe the world, then to say that an action is person-caused (or situation-caused) is just to express matching (or mismatching) noncognitive attitudes toward an agent and her action. On this view, the folk notion of self-disclosure is not an action-theoretic relation at all and self-disclosure judgments are not even in the business of psychological description.
are self-disclosing when building and testing normative theories.\textsuperscript{14} So, inasmuch as the arguments for the mismatch theory lower our confidence in those intuitive judgments, they should also lower our confidence in the normative theories we have constructed using them. Second, philosophers pursuing deep self accounts of moral responsibility appear to assume that they are, in part, engaged in a project of \textit{analysis}. In other words, they take it that their task is to articulate and refine a folk action-theoretic notion. But if the mismatch theory gives the correct account of the relevant folk psychology, there is simply no action-theoretic relation in the vicinity to analyze. Despite appearances, assertions like “Action A really belongs to agent S,” “Doing A reveals the real S,” and all the rest, do not refer to a naturalistic relation between agents and some of their actions. If this is right, philosophers must instead \textit{construct} an action-theoretic relation to play the role of self-disclosure. But now, as with any invented notion, they will have to explain why we should care whether an action is self-disclosing.

Attributional judgments are supposed to capture something about an agent’s psychology; they’re supposed to be the kind of thing that might serve as a descriptive basis for moral judgments. However, as I have argued, whether the cause of an action is perceived to lie more in agential or situational factors is not determined by anything of the sort that a naturalistic scientist might describe. Rather, the perceived location of the cause is significantly determined by moral attitudes. Second, I have argued that the social scientific person/situation distinction and the folk psychological notion of self-disclosure are really two windows looking onto the one bit of folk psychology. Third, if I am right about the psychology, intuitions about whether an action is self-disclosing do not track an action-theoretic relation and cannot serve as a non-moral basis for judgments about praise and blame. Judgments about whether an action is self-disclosing are thoroughly moral to begin with.

\textsuperscript{14}Or so it seems to many, including myself—but this is controversial, see, e.g., Williamson, 2008.
Chapter 3

Attribution, Determinism, Luck

3.1 The asymmetric threat of determinism

Philosophers sometimes observe that incompatibilism finds its strongest intuitive support in cases of particularly bad actions (e.g., Wolf, 1980; Nelkin, 2011). Reflecting on the determining causes of morally bad actions often lowers our confidence that agents are responsible. The following is a true story that I love for illustrating this point.¹

Hector Black fought in the second world war. He came home, went to Harvard, and there he met and married his wife, Susie. When he heard about the civil rights movement, he packed a bag, moved to the neighborhood where Martin Luther King lived, and asked King, “How can I help?” That’s where he met Patricia—a neighborhood kid who came over to his place all the time, “like a lot of kids, but she hung around.” Trish was a neglected child: her mom was an alcoholic who would often drink through the rent money. When she turned 11 she asked Hector if he would adopt her. He agreed, and she flourished. Eventually she went to college,

¹I first heard it on NPR’s Radiolab <http://www.radiolab.org/story/317421-blame/> , and my telling closely follows theirs.
returning home after earning her degree to work as a primary school teacher. She even took in a few neglected kids as Hector had done for her.

This is where the story turns dark. Ivan is addicted to crack. He burgles Patricia’s house looking for things he can sell or trade for drugs. He is in her bedroom when he hears Patricia come home. He hides in a wardrobe, planning to escape through the back window once the coast is clear. But Patricia opens the wardrobe. She falls backwards, and Ivan ties her up. “He says he wants sex. She says you’ll have to kill me first. So he does.”

At first, all Hector Black—Patricia’s father—can think of is revenge. He wants Ivan Simpson to suffer like Ivan had made Patricia suffer. But then he learns more about Ivan. He learns that Ivan was born in a mental hospital to a mother suffering from schizophrenia. When Ivan “was 11 years old, his mother took him and his brother and sister to a swimming pool and said God was ordering her to destroy them.” Ivan escaped, but not before watching his mother drown his little sister. The story actually gets worse, if you can imagine that.

After Hector learns Ivan’s terrible story he writes to the district attorney. He says he doesn’t want to pursue the death penalty. On the final day of Ivan’s hearing, Hector waits until all of the family members have read their statements, and then he gets up to read his. At one point, he says this: “I don’t know if I’ve forgiven you, Ivan Simpson, but I don’t hate you. I hate with all my soul what you did to my daughter.” Then somebody turns him around to face Ivan Simpson: “I was looking at him, in his eyes . . . And then tears were streaming down his cheeks. It was the first time I looked into his eyes, and it was like a soul in hell.”

If you listen to Hector talk about Ivan, it is clear that he does not just think Ivan isn’t fully responsible in the sense of accountability. He does not just think that it would be unfair to inflict (retributive) punishment on Ivan—perhaps because, given his terrible childhood, Ivan never had a fair shot at avoiding those sanctions. That’s
not it at all. Rather, Hector is less inclined to blame Ivan in the sense of *attributability*. Blame, in the sense of attributability, is about locating a moral fault in someone; it’s about judging a person’s true self. And even in the sense of attributability, Hector’s blame softened when he learned about Ivan’s childhood. To Hector, the moral fault, in some sense, did not belong to Ivan—it was not a part of his true self. The most remarkable thing about this story is that Hector and Ivan still exchange letters regularly. They’ve been doing so for years. Hector even sends Ivan a present on his birthday every year, which Ivan receives in his prison cell.

When we reflect on the causes of Ivan’s bad action, our inclination to blame him softens, just as Hector Black’s did. Even if we cannot say exactly what it is, Ivan seems to have a partial excuse; he is not fully responsible. Reflection on the determining causes of bad actions often has this powerful effect. And it doesn’t seem to be a matter of “there but for the grace of God go I.” If God knew Ivan’s story, it seems like it would be a mistake even for Him to blame Ivan fully.

It is a very interesting fact that parallel observations in the case of good actions often do not seem to threaten our sense that an agent is fully responsible—that is, that she is fully worthy of praise. Dwell as much as you like on the causal antecedents of his good actions—his virtue and his powerful moral commitments—and you will not lose your sense that Gandhi deserves praise for all his good deeds. Reflect on the details of his childhood, details that determined that he would do his morally great deeds, and I don’t think your inclination to praise Gandhi will be at all diminished.

---

2Learning about Ivan’s childhood tends to temper attributional and retributive impulses, but very few people come to think that he is *entirely* blameless (unaccountable for his actions). Rather, the conditions of Ivan’s childhood provide a *partial* excuse. For self-disclosure accounts of responsibility to capture this fact they must hold that something can belong to a person’s true self to a greater or lesser degree, or allow that self-disclosure alone does not determine our judgments of responsibility (or both). If the phenomena of folk attribution are a manifestation of psychological essentialism, the question is whether a disposition can be undergirded by a person’s essence to a greater or lesser degree. While psychologists tend to talk as if dispositions are either wholly essential or wholly superficial, to my knowledge, nothing in the empirical literature actually requires this assumption. It is also likely that responsibility (accountability) judgments are sensitive to features other than self-disclosure—beliefs about obligations and local norms, to take just two examples, might exert an independent influence on responsibility judgments.
To me, at least, the fact that his good deeds have determining causes—the fact that he may literally have had to do them—just seems utterly irrelevant to the question of whether he is praiseworthy. But that’s very odd. After all, metaphysically speaking, Gandhi’s and Ivan Simpson’s cases are very much alike: Gandhi and Ivan were equally determined agents. So why should reflection on the determining causes of their actions not have an equal effect on our sense that they are morally accountable for their actions? How could determinism threaten to undermine responsibility for bad actions and not responsibility for good actions?

3.1.1 The rational abilities approach

A general strategy for explaining this asymmetry holds that freedom in the sense required for responsibility is a matter of possessing the ability to do the right thing for the right reasons. Susan Wolf exemplifies this approach when she proposes that S did A freely if and only if S could have done otherwise had there been good and sufficient reason to do so (1980, p.212).

Thus, according to Wolf, when we assess whether an agent is accountable, we first check to see if there were good reasons for the agent to do otherwise. If we find that there were—if the agent was not doing what there was good reason to do—then we require that she could have done otherwise. We require this because if she was unable to do otherwise, then she did not have the ability to act for the right reasons. But if we find that there was no good reason to do otherwise—if the agent was doing what there was good reason to do, and was doing so for that very reason—then we are uninterested in whether she could have done otherwise. This makes intuitive sense, for it would seem absurd to insist that a reflective and integrated agent who must do the right thing for the very reasons that make it right is less deserving of praise than an agent who does the right thing but could have chosen not to.\(^3\)

\(^3\)On Wolf’s view, for a psychologically determined agent to be responsible for her good deeds, it
To take the bad case first, consider Ivan. When he kills Patricia he is not doing what there is good and sufficient reason to do. So to assess whether he is blameworthy we need to assess a counterfactual; for we need to know whether he was able to do the right thing for the right reasons, and the abuse he suffered as a child might makes us doubt that he was. His childhood meant that he was never able to develop his moral faculties, and this diminishes our sense that he is responsible. On the other hand, when we ask whether Gandhi is praiseworthy for his actions—are they attributable to him? does he deserve credit for them?—we suppose that he was indeed doing what there was good and sufficient reason to do, and moreover, he was doing so for precisely those reasons. So we are uninterested in whether he was able to do otherwise. If his actions resulted from a properly functioning moral faculty, then we do not mind that he had to perform them. This is why, on the rational abilities view, determinism threatens our sense that an agent is blameworthy, but is irrelevant to questions concerning whether she is praiseworthy.

Pizarro, Uhlmann and Salovey (2003) explored a similar asymmetry experimentally and their findings appear to support the rational abilities view (inasmuch as it is understood as a claim in moral psychology). Consider the following vignettes:

**Deliberate:** Jack calmly and deliberately smashed the window of the car parked in front of him, because it was parked too close to his.

**Impulsive:** Because of his overwhelming and uncontrollable anger, Jack impulsively smashed the window of the car parked in front of him because it was parked too close to his.

must not seem as if she has simply been “programmed” to do the right thing. Rather, she must be determined “in the right way.” The values which, in concert with her beliefs, determine her actions, should cohere with one another and mesh with her “ideals of how one ought to live.” The idea may be that the agent must be determined such that her good actions are self-disclosing—the attitudes which motivate her actions must belong to her true self.
As you might expect, Pizarro et al. found that people assign less blame to Jack when his action results from overwhelming emotion than they do when his action results from cool-headed deliberation. Plausibly, this is because people suspect that Jack’s overwhelming and uncontrollable anger masks his ability to act on the right reasons. But now consider the following pair of cases:

**Deliberate:** Jack deliberately and intentionally gave the homeless man his only jacket, even though it was freezing outside.

**Impulsive:** Because of his overwhelming and uncontrollable sympathy, Jack impulsively gave the homeless man his only jacket, even though it was freezing outside.

In this case, the agents were regarded as equally praiseworthy: when assigning praise, people did not draw a distinction between deliberate and impulsive actions. So it seems that acting from an overwhelming emotion tends to reduce our sense that an agent is morally blameworthy, but not that he is morally praiseworthy.

The rational abilities view explains this asymmetry. When we read about Jack smashing in a car window, we judge that he is not doing what there is good reason to do. So we are interested to know whether he could have done otherwise. We suppose that normal adults are able to do otherwise, unless something unusual is masking this ability, and when Jack acts calmly and deliberately we suppose that there is nothing masking his ability to do otherwise. So we judge that he is fully blameworthy. But when he acts from overwhelming and uncontrollable anger we begin to suspect that he did not, at that moment, have the ability to act for the right reasons. So he has an excuse; he is not fully blameworthy.

Things look different when we consider the positive case. When we read about Jack giving his coat to the homeless man we judge that he is doing the right thing for the very reasons that make it right. So there is no question about whether he possess
that ability; so we are uninterested in whether he could have done otherwise; so he is fully praiseworthy, whether or not he had to act as he did. (My own sense is that he is more praiseworthy when he gives his coat away out of uncontrollable sympathy than out of cool-headed deliberation.)

3.1.2 The mismatch approach

The mismatch theory offers a competing explanation for these results, and for the asymmetry in our intuitions about Ivan and Gandhi. For if we begin with the assumption that all of these people are, deep down, morally good, then we will be more inclined to attribute their actions to non-agential forces when they act badly. When Jack smashes in the car window we perceive a mismatch in the moral valence of the actor and his action, and this inclines us to attribute the action to something outside of Jack’s true self. Since we believe that Jack’s essence is good, his uncontrollable and overwhelming anger seems to lie outside of his true self, and therefore provides an ideal attribution for his actions. So his action does not appear self-disclosing, and this is what leads us to attribute less blame.4

4Doesn’t the mismatch theory predict that Jack is blameless even when he calmly and deliberately smashes in the window? After all, we presume that at his very essence Jack is morally good, so even in the cool-headed case, shouldn’t we perceive a blame-eliminating mismatch? No; the mismatch theory only predicts (correctly—see Newman et al., 2014) that people will be even more inclined to blame Jack when they believe he has an evil true self. So compared to the case of Evil Jack, blame is somewhat suppressed, although not eliminated, when Good Jack calmly smashes in the window. Moreover, people tend to be more conservative with blame than they are with praise—they praise more readily and more intensely than they blame—suggesting that in the ordinary case blame is suppressed from where it would be were there no mismatch. (Of course it must be admitted that there are several other plausible putative explanations for this fact; for instance, the costs of undeserved blame are typically higher than those of undeserved praise (Watson, 1996).)

It is also no part of the mismatch theory that we never revise our beliefs about a person’s true self. Indeed, some action types seem especially good at causing us to revise, at least to some degree. As Leslie (in press) observes, while worrying on the odd occasion does not make one a worrier, murdering on the odd occasion certainly makes one a murderer (also see Mende-Siedlecki et al., 2013). Thus our presumption that a person is fundamentally morally good can be overridden, at least to some degree, and the mismatch theory’s prediction is that it is more likely to be overridden when there is no situational attribution for that person’s bad actions. That is why our initial reaction to learning about Ivan’s actions is extreme blame; it is only after we learn his story—that situational attributions have been made available—that we are again able to see his true self as morally good.
In the positive case we see Jack, a morally good person, doing a morally good action, and this inclines us to see his action as arising out of his true self. In this case his overwhelming and uncontrollable emotions (sympathy for the homeless man) are positive, so we see them as lying within his true self. So the action which results from these feelings seems fully self-disclosing, and consequently Jack seems no less praiseworthy than when he gives the man his coat out of cool-headed deliberation.

A similar process may explain Hector Black’s reaction to learning about Ivan Simpson’s story. Learning about Ivan’s history makes situational attributions available—it makes it easier to see his actions as arising from factors outside of his true self. The perceived mismatch between Ivan and his action inclines us to make use of that easy possibility, and this is what tempers our sense that he is blameworthy. Of course, learning about Gandhi’s story also makes situational explanations for his actions available, but since there is no mismatch between Gandhi’s moral valence and the valence of his actions, we are not very inclined to explain his actions in situational terms. So we continue to see them as fully self-disclosing, and hence we see Gandhi as fully responsible.

3.1.3 Where the two approaches differ

Both the rational abilities view and the mismatch theory model the asymmetry in our intuitions in the case of morally good agents. But they make very different predictions when it comes to fundamentally morally bad agents.\(^5\) For the ratio-

\(^5\)If the mismatch theory’s explanation is correct, individual differences that influence attributional biases should also influence the asymmetry in our intuitions about determined agents. Effect sizes for self-serving bias—the tendency to disproportionately attribute good events to internal causes—are on average three times smaller in Asians (d = .3) than North Americans (d = 1.05), and more than twice as small in samples with psychopathology (depression, anxiety, or attention-deficit/hyperactivity disorder) (d = .48) than without (d = 1.28) (Mezulis et al., 2004).

To take a concrete example, Asians should be less inclined than North Americans to provide internal attributions for Jack’s giving his coat to the homeless man. So, while Asians should still be more inclined to attribute Jack’s giving his coat away to internal causes than his smashing the car window, the asymmetry should be somewhat reduced, compared to North Americans. Thus, the
nal abilities view seems indifferent to agents’ moral valence, so it predicts that the asymmetry in our intuitions noted above will appear both when considering morally good agents and when considering evil agents. The mismatch theory, on the other hand, predicts that inasmuch as we think of an agent as morally bad, the asymmetry will be reduced, since we will perceive less of a mismatch between her bad actions and her true self. So consider the first pair of cases again:

Deliberate: Jack calmly and deliberately smashed the window of the car parked in front of him, because it was parked too close to his.

Impulsive: Because of his overwhelming and uncontrollable anger, Jack impulsively smashed the window of the car parked in front of him, because it was parked too close to his.

The mismatch theory predicts that, inasmuch as we can think of Jack as morally evil, the presence of uncontrollable anger should not incline us towards external attributions for his bad actions. For inasmuch as we believe that deep down he is morally evil, his negative emotions will appear to lie within his true self, and this will lead us to see his impulsively smashing the car window out of anger as self-disclosing. So we should blame him to the same degree when he smashes the window impulsively as we do when he does so deliberately.

When I try to imagine that Jack is morally evil at his very core, the asymmetry in my intuitions seems to disappear. Admittedly, this judgment is a subtle one, so it is a good thing that just this case was tested by Newman, Freitas, and Knobe (2014). They tested Pizarro et al.’s window smashing prompts, but primed subjects to think of the agent as either fundamentally morally good or evil. When subjects were primed to think of Jack as morally good, Newman et al. replicated Pizarro et al.’s mismatch explanation predicts that the determinism asymmetry will be attenuated for Asians (and people with various psychopathologies). Testing this prediction is a task for future work; I mention it here only as another example of where the mismatch and rational abilities accounts come apart.
results: subjects reported reduced blame when Jack acted out of overwhelming and uncontrollable anger. But when subjects were primed to think of Jack as morally evil, impulsiveness did not lead them to discount blame (Cohen’s $d = .35$).

Newman et al. hypothesize that the difference is due to whether participants see Jack’s emotional state as internal or external to his true self, and the mismatch theory helps to complete this picture. In both cases, Jack’s smashing the window is thought to express his overwhelming and uncontrollable anger. But whereas his anger provides a situation-attribution when Jack is described as morally good, the very same state provides a person-attribution when he is described as morally bad. So Evil Jack’s impulsive action appears more self-disclosing than Good Jack’s, and that is why we judge that the bad agent who acts impulsively is more responsible than the good agent who does the same.

It may be objected that the rational abilities view can make sense of this pattern, since evil agents might be thought to be responsible for their characters. Perhaps we suppose that Evil Jack is unable to act for the right reasons, but his responsibility for his actions *traces* back to a time when he did meet the conditions of responsibility. That is, perhaps we think about some evil agents in roughly the way we might think about some severely intoxicated agents. If Sarah goes on a murderous rampage after intentionally taking a drug that she knows will make her uncontrollably violent, she may be responsible for her actions while intoxicated, even though at the time of action she did not have the capacity to do the right thing for the right reasons. Her responsibility at the time of action might trace back to a time when she did meet the conditions for responsibility; that is, to when she decided to take the drug. But even putting aside serious difficulties with the idea of tracing (see, e.g., Vargas, 2005), this is not a plausible explanation for Newman et al.’s findings. For they primed participants to think that Jack’s evil traits had been stable all his life:

Ever since [Jack] was born, it was clear that there was something dis-
tinctive about his personality. He sometimes did good things to other people, but deep down in his very essence, he was a fundamentally evil person. At the very core of his being, he had no compassion for other people and no concern at all about their well-being.

Participants were explicitly told that there is no time in Jack’s history where he was morally different to how he is at the time of action. So it is hard to see how they could trace his responsibility for smashing the window back to a time when he was responsible for acquiring his moral character. So it is hard to see how the rational abilities view can explain why overwhelming and uncontrollable anger tempers our inclination to blame Good Jack for his action but does not temper our inclination to blame Evil Jack for his. On the other hand, this pattern is exactly what we would expect to see if the mismatch theory were true.

According to the mismatch theory, the feature of these cases that modulates our inclination to blame is not that Jack is unable to act for the right reasons. Rather, his uncontrollable anger both renders him unable to act for the right reasons and provides a situational attribution for his bad action. According to the mismatch theory, it is the ready availability of this situational attribution, in concert with the contrasting valences of Jack and his action, that tempers our inclination to blame.

Thus we are led to a radically surprising prediction: it should be possible to find the same asymmetry in cases where the agents are able to act for the right reasons. In fact, it should be possible to find the same asymmetry even in cases where the agents meet just about any plausible conditions for free action. In the following section, I will argue that many such cases exist: luck can have the same asymmetrical effect on our intuitions, even when there is no question about whether the agents were able to act for the right reasons.
3.2 Luck and responsibility

Reflecting on the role that luck played in bringing about a bad action often diminishes our sense that the agent is responsible. When we reflect on the fact that it was just a matter of bad luck that the boy was born in Germany toward the beginning of the twentieth century, and that had he been born in Australia he would probably have lived a morally innocuous life rather than becoming a Nazi officer, it tends to shake our sense that he is fully responsible. Which opportunities for action we find arrayed around us is, to a large extent, a matter of luck; and what we do is, in part, a matter of what opportunities for action we face (Nagel, 1979; Williams, 1981).

Interestingly, reflecting on the role of luck in bringing about virtuous action does not seem to have the same effect. So far as Bill Gates was concerned, it was sheer luck that the Mothers’ Club of his preparatory school decided to buy a Teletype terminal and access to a GE computer for the school’s students. Had this chancy event or countless others over which Gates had no control not occurred, he would not have founded Microsoft. So it was, to a considerable extent, a matter of luck that Gates accumulated his fortune and funded important humanitarian projects through the Bill and Melinda Gates Foundation. But the role that luck has played in Gates’ contributions to education, to treating and preventing HIV, to improving sanitation and access to water in developing countries … seems utterly irrelevant to

---

6The kind of luck philosophers (and psychologists) have focused on most is resultant luck (Nagel, 1979)—or luck in how one’s actions turn out. The classic case of resultant luck: a child darts out into the path of a lorry driver who has negligently failed to have his breaks checked; he cannot stop and consequently kills the child. This is a case of bad resultant luck inasmuch as we would correctly judge the driver less harshly had the child not darted into his path (he would deserve a fine and not several years in prison). In Nagel’s classification, the cases just mentioned in the text, and those examined below, are all cases of circumstantial luck—luck in the circumstances in which one finds oneself. However, because the circumstances of interest are the circumstances of the agent’s early life (where she is born and raised, whether she is adopted, etc.), these cases are also indirectly cases of constitutive luck—or luck in the kind of person one is. In this chapter I leave the phenomena of resultant luck to the side. Looking for the asymmetry with cases of resultant luck, and separating out the effects of constitutive luck, are tasks for future research. (However, I strongly suspect that the asymmetry will appear with cases of any kind of luck).
whether he deserves credit for his remarkable deeds. So it seems that luck reduces our sense that agents are responsible for their bad actions far more readily than it reduces our sense that they are responsible for their good deeds. The following section describes an experiment designed to explore this asymmetry. I return to consider the best explanation for the determinism asymmetry in Section 3.2.3.

3.2.1 Method

To test the hypothesis that luck would asymmetrically affect responsibility judgments, 250 participants were recruited via mTurk (M_{age} 30.1; SD_{age} = 8.8; 42% female) to participate in a 2 (GOOD vs BAD) x 2 (LUCK vs NO LUCK) between subjects study.

Participants read one of four vignettes about a person, born in the United States in the early 1800s, who does something morally bad (owns and mistreat slaves) or morally good (helps African Americans escape from slavery). In two conditions, the vignette emphasized the role of chance in the agent’s life (I used the word *chance* because *luck* tends to be associated with good things):^7

**Bad luck**: Chance played a big role in Bill’s life. He was born in the Northern United States in the early 1800s, but as a baby he was adopted by Southern plantation owners who raised him in the South. If he had been raised by his biological parents, he would have grown up in the North, and he would have led a morally better life. But as a matter of fact, Bill himself went on to own slaves, and in this way he hurt many people over the course of his life.

**Good luck**: Chance played a big role in Bill’s life. He was born in the Southern United States in the early 1800s, but as a baby he was adopted by Northern abolitionists who raised him in the North. If he had been

---

^7To test the robustness of this effect, and to help mitigate the possible effects of survey design, I tested two different wordings of these vignettes. The same effect emerged in both versions.
raised by his biological parents, he would have grown up in the South, and he would have led a morally worse life. But as a matter of fact, Bill himself went on to work on the Underground Railroad to help people escape from slavery, and in this way he helped many people over the course of his life.

The control conditions simply omitted the features expected to focus participants on the role of luck in the agents’ lives:

Control: Bill was born in the Southern [Northern] United States in the early 1800s. Bill owned many slaves [worked on the Underground Railroad to help people escape from slavery] and in this way he hurt [helped] many people over the course of his life.

Following each vignette participants responded to two questions: “How negatively [positively] does Bill deserve to be judged?” (1 = Not at all negatively [positively], 7 = Extremely negatively [positively]); “How much blame [praise] does Bill deserve?” (1 = No blame [praise] at all, 7 = Extreme blame [praise]). These scales were highly consistent (Cronbach’s $\alpha = .91$), so I averaged scores to create a composite scale.

Combining scales for responsibility and self-disclosure into the one survey appeared to confuse subjects, so I recruited a different sample (N = 250, $M_{\text{age}}$ 36.4; $SD_{\text{age}}$ = 10.1; 38% female) to rate agreement with the statement, “Helping [harming] people did not reflect Bill’s true self—the person he really is deep down” (1 = strongly disagree, 9 = strongly agree). Responses to this scale are inverted to make larger values indicate higher self-disclosure ratings. All items were counterbalanced for order, and followed by a screening question to detect participants who were not reading. (Participants who answered this question incorrectly were excluded from the study; for discussion, see Downs et al. 2010; Goodman et al. 2013.)
3.2.2 Results

An ANOVA test confirmed that the interaction between valence and luck on responsibility judgments is statistically significant ($p = .01$). As predicted, across the **BAD** conditions, where Bill is adopted by Southern plantation owners, drawing participants’ attention to the role that luck played in Bill’s becoming a slave owner led to a large reduction in responsibility ratings compared to the no-luck condition (Cohen’s $d = .64$; $M_{\text{diff}} = 0.93$ [0.4 - 1.47]; $p = .001$). However, luck did not meaningfully affect responsibility ratings across the **GOOD** conditions, where Bill becomes a worker on the Underground Railroad (Cohen’s $d = .19$; $M_{\text{diff}} = 0.14$ [-.13 - .41]; $p = .32$). So the interesting result we see here is that reflecting on the presence of luck appears to reduce people’s sense that Northern-born Bill is responsible for his bad deeds, but has little or no effect on their sense that Southern-born Bill is responsible for his good deeds (see Figure 3.1 on the following page).

Similarly, responses to the self-disclosure question patterned as predicted. An ANOVA test confirmed that the interaction between valence and luck on self-disclosure judgments is statistically significant ($p = .002$). Across the **BAD** conditions, luck had a meaningful effect on whether participants perceived Bill’s owning and hurting slaves as revealing of his true self (Cohen’s $d = .57$; $M_{\text{diff}} = 1.2$ [0.6 - 1.7]; $p < .001$). But across the **GOOD** conditions, emphasizing the role of luck had no effect whatsoever on whether participants perceived Bill’s helping people to escape from slavery as revealing of his true self (Cohen’s $d = .01$; $M_{\text{diff}} = .01$ [-0.54 - 0.56]; $p = .97$). In other words, while telling participants that it was a matter of luck that Bill led a morally **bad** life meaningfully reduced self-disclosure and blame ratings—just as philosophers would predict—telling participants that it was a matter of luck that Bill led a morally **good** life did not stop them seeing his actions as both highly self-disclosing and highly praiseworthy (see Figure 3.2 on the next page).
**Figure 3.1:** The interaction of luck and valence on responsibility ratings $N = 242$.

Valence * Luck: $p = .01$, partial $\eta^2 = .026$; Valence: $p < .001$, partial $\eta^2 = .233$; Luck: $p = .001$, partial $\eta^2 = .047$.

**Figure 3.2:** The interaction of luck and valence on self-disclosure ratings. $N = 460$ (combining data from a replication study). Valence * Luck: $p = .002$, partial $\eta^2 = .021$; Valence: $p < .001$, partial $\eta^2 = .121$; Luck: $p = .002$, partial $\eta^2 = .020$. 
Finally, participants consistently rated the morally good cases as more self-disclosing than the morally bad cases, both in the LUCK conditions (Cohen’s $d = 1.1; p < .001$) and the NO LUCK conditions (Cohen’s $d = 0.44; p = .002$), providing some further evidence that people believe implicitly that there’s something morally good deep down inside of everyone.

### 3.2.3 Discussion

**The determinism asymmetry**

The asymmetry revealed in the present study is very similar to what we find in our intuitions about the threat of determinism. But what’s especially interesting about the luck asymmetry is that it arises even when there is no suggestion that agents are in any way unable to do otherwise or that they lack the capacity to do the right thing for the right reasons. Remarkably, the asymmetry appears even when there is no suggestion that the agents were overcome by irresistible emotions or that they were severely abused as children. So whatever explains the asymmetry in this case, it is not plausibly that we only require that agents could have done otherwise had there been good reason to do so. On the other hand, the mismatch explanation applies equally well: just as considering the causal antecedents of an action can focus our attention on situational explanations for the action, so too can considering the role that luck played in its genesis. In each case these situational explanations make it possible to see the action as less revealing of the agent’s true self, and this, in concert with a perceived mismatch in the valence of the agent and her action, may be what accounts for our sense that the agent is less blameworthy.

While the hypothesis that people believe responsible agents must have the ability to do the right thing for the right reasons predicts this asymmetry in the case of good agents, it makes the wrong predictions in the case of bad agents, and it does
not apply to the luck asymmetry demonstrated here. Since these two asymmetries appear to be closely related—if not identical—other things equal, we should prefer an account that can explain our judgments in both cases. So we should prefer the mismatch explanation.

**Watson’s ambiguity explanation**

These results also help to address an alternative explanation for the determinism asymmetry proposed by Watson (1996), namely, that the asymmetry arises from a tacit shift between two conceptions of responsibility.

On Watson’s view, when we judge that a determined agent is responsible for her good deeds, we are making a judgment about responsibility in the sense of *attributability*. The content of this judgment is roughly: the action reflects the agent’s true self. But on Watson’s view, when we judge that a determined agent is not responsible for her bad deeds, we are making a judgment about responsibility in the sense of *accountability*. And the content of this judgment does not concern whether the action belonged to the agent’s true self; rather, it concerns whether the agent deserves special treatment just on account of her action. According to Watson, from within each perspective there is no asymmetry: if their actions are determined, both agents are responsible in the sense of attributability, and neither is responsible in the sense of accountability.

At the beginning of this chapter I suggested that learning about the horrible abuse that Ivan Simpson suffered as a child appears to reduce his responsibility in both senses—learning about his childhood naturally leads both to the thought that the moral fault did not entirely belong to his true self, and to the thought that he did not deserve the most extreme possible punishment for his crime. (It is very tempting to think that he did not deserve such extreme punishment precisely because the moral fault did not belong to him.) If that is right, then Ivan’s case presents a
problem for the ambiguity explanation of the asymmetry in our use of the word ‘responsible’, and the results described in this section seem to add further weight to this problem. For as we have seen, the (closely related) luck asymmetry seems to arise both when we ask explicitly about attributability and when we ask explicitly about accountability. If ordinary people actually distinguish these putative senses of responsibility, they do not seem to do so in the way Watson suggests.\footnote{Common tests suggest this is unlikely to be a simple case of lexical ambiguity (Zwicky and Sadock, 1975). It is not difficult to think of cases where it would seem felicitous to say, of two determined agents, “A is responsible and so is B,” where A performs a good action and B performs a bad action. However, much as one cannot say, “A went to the bank and so did B,” if A went to the side of the river and B went to the financial institution, the sentence “A is responsible and so is B” would not be felicitous in this context if its truth required A to be responsible in the sense of attributability and B to be responsible in the sense of accountability.}

**Luck and the control principle**

Many philosophers have thought that our responsibility is limited to those things that are under our control. Luck can shake our sense that an agent is responsible because (it is thought) we believe that “one cannot be more culpable or estimable for anything than one is for that fraction of it which is under one’s control” (Nagel, 1979, p. 451). Applied to the cases above, this principle suggests that we discount blame because we recognize that Bill’s owning and mistreating his slaves is, to some degree, a function of his being adopted by slave-owning Southerners—and that was not something that was ever under his control.

However, if the results I have just described can be taken at face value, this cannot be right. For this hypothesis predicts that learning about the equally large role that luck played in Bill’s coming to work on the Underground Railroad should reduce our sense that Bill is praiseworthy. But as I have argued, and as the results presented here suggest, this simply isn’t true.
CHAPTER 3. ATTRIBUTION, DETERMINISM, LUCK

Luck and modality

When philosophers analyze the concept of luck, they often attempt to do so in more or less purely modal terms (see, e.g., Coffman, 2007; Levy, 2011; Pritchard, 2006). The rough idea common to these analyses is that an event is lucky just in case it does not occur in sufficiently many nearby possible worlds in which its initial conditions are held fixed. Philosophers disagree about what counts as sufficiently many worlds, about how to identify the relevant initial conditions, about whether insignificant events can be lucky, and so forth. However, they usually agree that the relevant notion of luck is fundamentally modal and evaluatively insensitive.

The present results suggest that, inasmuch as these philosophers hope to analyze a folk concept that bears on people’s sense that agents are responsible, this understanding is fundamentally deficient. Whether circumstantial luck tempers our sense that an agent is responsible does not depend solely on whether she finds herself in the relevant circumstances in many nearby possible worlds; rather, the modal and evaluative features of her case seem to interact when influencing our judgments of responsibility.

Relation to other research

Some previous research suggests that, compared to morally good actions, morally bad actions tend to have an outsized effect on how people judge an agent’s moral valence (Reeder and Coover, 1986; Skowronski and Carlston, 1987). Moreover, some research finds that the effect of bad behavior on impression formation tends to be not only larger than the effect of good behavior, but considerably more impervious to mitigating features of the agents’ circumstances (Reeder and Spores, 1983). According to this research, if I see you steal something, for example, I’m likely to infer that you’re a bad person, without much regard to the situation you were in—for example, whether someone was urging you to steal. On the other
hand, if I see you donate to charity, I will only infer that you’re a good person if there were no situational pressures “facilitating” your good action (ibid.).

Several putative explanations for these effects have been offered. One holds that bad behavior is believed to be statistically less common than good behavior, and is therefore more informative for the purpose of categorization (Fiske, 1980; Mende-Siedlecki et al., 2013). Another related explanation begins with the thought that, while it is difficult to induce a good person to behave immorally, a bad person may be expected to behave morally whenever she believes that doing so serves her narrow self-interest. So, when you observe someone acting immorally, you can be pretty confident that she is indeed a bad person; but you cannot be so sure that someone is behaving morally because she is a good person—after all, she might be a bad person who is doing the right thing for her own expedient reasons.

The experimental results described here seem to reveal a radically different side of attribution and social perception. Indeed, these results are the opposite of what we would expect on the basis of the research just mentioned: distant situational attributions (Bill’s adoption) seem to have a large effect on how people explain bad actions (owning and mistreating slaves), but this effect actually disappears when they explain morally good actions (working on the underground railroad).

Results like those reported in Pizarro et al. (2003) are also in tension with previous research. The presence of overwhelming emotion affects attributions for bad actions (smashing in a car window), but has no effect on attributions for good actions (giving one’s jacket to a homeless man), as long as the actor is thought to have a morally good essence (Newman et al., 2014). Similarly, the results reported in Newman et al. (2013), and in the second chapter of this dissertation, seem to involve participants judging that people are, at their very essences, morally good, even when these people are behaving in a way that participants find extremely objectionable (Mark hitting on another man; Kate deciding to have an abortion).
However, the results described here do align with other, independent bodies of evidence. Much research attests to the fact that we have a strong bias towards evaluating people favorably (e.g., Hallahan et al., 1997; Klar and Giladi, 1997; Sears, 1983). This research does not distinguish among levels of the self, but it is reasonable to suppose that people’s beliefs about the true self will typically inform their judgments of others’ valences, their inclinations to provide person-attributions for others’ successes and good deeds, and so on.

Recently, researchers have suggested that these positivity biases are connected to our psychological essentialism (Newman et al., 2013, 2014). Evidence is accumulating for the idea that we tend to think of essences as valenced, both in the cases of biological categories (Lockhart et al., 2002) and non-biological categories (Knobe et al., 2013; De Freitas et al., 2014). This research suggests that when we look to see who a person really is deep down, we will tend to find whatever we think of as morally good. Against this background, it is far less surprising that the availability of situational attributions should lead us to discount bad actions when making judgments about responsibility and self-disclosure.

Better understanding how these seemingly opposed pictures of attribution and social perception might be reconciled is an important topic for future research.

### 3.3 Conclusion

The mismatch explanation for the asymmetric threat of determinism is that we focus on different causes, or locate the same cause in different places, depending on whether the moral valence of the action matches that of the agent. When there is a mismatch in valence, the cause is seen more as arising from outside of the agent’s true self—the action is not self-disclosing—and this reduces our sense that the agent is responsible. But when there is a match, we can focus on different causes or see
the same cause as lying within the agent’s true self, thereby leaving our sense that
the agent is responsible fully in tact. Given that we tend to think of agents as having
morally good essences, this explains why reflecting on the causes of an action tends
to reduce our sense that agents are blameworthy more readily than it reduces our
sense that they are praiseworthy.

Philosophers have provided competing accounts for this asymmetry in our
intuitions about the conditions for moral responsibility, but none of these accounts
explains the full diversity of the phenomena. They do not explain why the asym-
metry is, as Newman et al.’s (2013) results suggest, sensitive to the perceived moral
valence of the agent. And they do not explain why the asymmetry appears to arise
even where there is no question about whether the action is determined, as the
experiments on luck presented here suggest. On the other hand, the mismatch
theory seems to be off to a good start: it can explain both why the asymmetry is
sensitive to the perceived valence of the agent, and why the asymmetry can come
completely apart from considerations about whether the action was determined.

Stepping back, the evidence suggests that Wolf (1980) was importantly on the
right track when she suggested that the applicability conditions for our concept
of responsibility include a normative component. But the way that our normative
beliefs get into the picture is not via considerations of whether the agent was doing
the right thing for the right reasons. Rather, the radically different account that
emerges here is that normative considerations influence our judgments because they
deploy influence how we explain actions. Normative considerations can incline us
toward situational attributions for an action, towards seeing it not as self-disclosing.
The role of determinism and luck is then to make such attributions readily available.

In saying this, I do not mean to suggest that we actually should hold people
responsible according to these patterns, but merely that they do play an important
role in our judgments of responsibility. For all I have said here, it may be that any
philosophical theory is right when it comes to the question of when we ought to hold people responsible (in any sense of the word), or when people are, in a metaphysical sense, really responsible. But I think the results presented here should still interest philosophers working on these normative and metaphysical questions. For correctly modeling people’s ordinary judgments only provides support for a philosophical view inasmuch as the explanation for why people have those judgments involves the truth of that view. Since the mismatch theory’s explanation for the determinism asymmetry does not involve the truth of any normative or metaphysical view of the conditions of responsibility, belief in the mismatch theory tends to undermine the support such views might otherwise have enjoyed from modeling people’s ordinary judgments.

For my part, however, the main interest of these results does not really derive from how they interact with normative or metaphysical theorizing. Rather, the question of how people actually think about responsibility is just extremely interesting in itself.
Bibliography


Chapter 4

Appendices

4.1 Reasons and Causes

We seem to have two broad approaches to explaining intentional actions. To a rough approximation, one involves citing the reasons in light of which an agent acts and the other involves citing the historical causes of those reasons. Causal histories help to explain why an agent has certain reasons, but are not themselves among her reasons.\(^1\) Whereas a reason explanation implies that an agent is aware of a certain content (a belief or desire content), she may be entirely unaware of the background causes of her reasons. Malle (2011) notes two features of causal history of reasons (CHR) explanations which can be attractive:

First, CHR explanations do not rely on an act of perspective taking. Therefore, any observer who cannot or does not want to take the agent’s subjective perspective to offer a reason explanation can still offer a CHR explanation (e.g., “Why were they joking at a bum’s expense?”—“I cannot explain why they did it other than they were intoxicated”). Second,

---

\(^1\)This may sound odd to philosophical ears. As is common in psychology, Malle (2011) uses ‘reason’ to refer to the beliefs and desires in light of which agents act (not to the facts that an agent may take to justify her action). CHR explanations should be understood as providing information about the causes of an agent’s beliefs and desires.
CHR explanations divert attention away from the agent’s reasoning and choice capacity [and so] portray the agent as less rational than reason explanations do (e.g., “When the medical people came by she was telling them what to do cause she was in panic”) (p.319).

This suggests an alternative explanation of the results reported above. In both studies, the internal attributions cited the agents’ reasons: Mark wanted to sleep with the man; Kate believed that it wasn’t the right time for her to start a family. The situational attributions did not: Mark hardly thought to himself, “I’m feeling tired, so I should hit on the man,” and Kate hardly thought to herself, “It’s easy for me to get an abortion at my college’s health clinic, so I should get an abortion.” These factors may help to explain why they had whatever reasons led them to act, but they are not plausibly among those reasons. Now perhaps subjects with high ATLG scores are more inclined to explain Mark’s action in terms of his being wiped out from teaching because (1), of the options on offer, only that explanation doesn’t require taking Mark’s perspective and (2), that explanation tends to present Mark’s action in a less rational light. After all, compared to subjects with low ALTG scores, high-scoring subjects will presumably be more averse to seriously considering Mark’s perspective and more likely to think his action irrational.

This hypothesis, however, cannot explain many of the results we have examined so far. It does not explain, for example, why the actor-observer asymmetry appears when actors and observers explain morally bad actions, and reverses when they explain morally good actions. Nor does it explain why intergroup attributions pattern as they do. It cannot explain these striking results because the internal/situational distinction and the reasons/causes distinction crosscut one another. Some intuitively internal attributions, like those in the studies reported here, do cite reasons; but equally, some cite causes. For example, “Mary made a
donation because she is kindhearted” is an intuitively internal attribution, but the fact that Mary is kindhearted is not plausibly among her reasons for donating.

In short, the patterning of internal and situational attributions cannot be described in terms of the reasons/causes dichotomy, so no hypothesis couched purely in terms of that dichotomy can explain the overall patterning of attributions.

4.2 Dissonance

Highly prejudiced people may find it difficult to believe that one man could ever want to sleep with another. Or perhaps people’s moral attitudes toward homosexuality influence their estimates of how common such desires are. In either case, we might expect that highly prejudiced people will avoid attributing desires of this kind. We should therefore expect these people to place comparably greater importance in the situational attributions, not so much because they are situational, but because they provide an alternative to positing erotic desires for the same sex. The same style of explanation may apply to the results of Study 2.

This hypothesis certainly sounds plausible, but I don’t think it is correct. The mismatch theory predicts that people with high ATLG scores will avoid internal attributions only to the extent that they believe that Mark is a good person. If instead they believe that Mark is a bad person, they should be comparably more willing to attribute to him an erotic desire for the same sex. The dissonance hypothesis, on the other hand, predicts that this willingness should be relatively insensitive to whether Mark is perceived as morally good or bad.

Although it is too early to say with confidence, preliminary experimental results appear to support the mismatch account. It should also be noted that the dissonance hypothesis does little to explain the patterning of self-disclosure judgments, actor-observer effects, or intergroup attributions.
4.3 Analyses using alternative exclusion criteria

Subjects in the original empirical studies reported here were recruited using Amazon’s Mechanical Turk online crowdsourcing marketplace. MTurk provides an affordable and convenient way to recruit subjects, but it also presents some problems. Apart from sampling biases, to my mind one of the most serious is that many MTurk workers are, effectively, professional survey takers, and making a living wage on MTurk requires you to work fast. This can show up as noise in data sourced using MTurk.

A pilot study with motivated undergraduates showed that people require about three minutes to read and respond to the surveys ‘briskly but thoroughly.’ Table 4.1 on page 110 shows that at least half of the online sample took one minute less than this. The patterning of responses to the ATLG items strongly suggests that these subjects were not contributing meaningful data. There are five items on the ATLG. For three of them, selecting strongly agree indicates a highly negative attitude to homosexuality; for the other two, it indicates a positive attitude. These last two items are therefore ‘reverse coded’—selecting strongly agree adds only one point to a subject’s score, rather than the usual five. One fast way of generating a score of 17 is therefore to select strongly agree for all five items (5 + 5 + 5 + 1 + 1). This unlikely response pattern was recorded over thirty times in Study 1. I conclude that these subjects were probably just clicking through the survey as fast as possible. This is further supported by the fact that the ATLG has a very high internal consistency (Cronbach’s $\alpha \approx 0.9$), nearly all subjects who scored 17 completed the survey in under 2:30 (see Figure 4.1), and the predicted relation of ATLG scores and situational attributions is considerably clearer when subjects who completed the survey quickly are excluded (see Figures 4.2 and 4.3).

The analyses presented in the main text relied on pilot studies to generate time-
Table 4.1: Descriptive statistics for survey completion times.

<table>
<thead>
<tr>
<th></th>
<th>Study 1 (Mark)</th>
<th>Study 2 (Kate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>168</td>
<td>158</td>
</tr>
<tr>
<td>Mean</td>
<td>2:29</td>
<td>2:20</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1:04</td>
<td>1:17</td>
</tr>
<tr>
<td>Minimum</td>
<td>0:33</td>
<td>0:32</td>
</tr>
<tr>
<td>Maximum</td>
<td>6:12</td>
<td>8:55</td>
</tr>
<tr>
<td>25th percentile</td>
<td>1:45</td>
<td>1:34</td>
</tr>
<tr>
<td>50th percentile</td>
<td>2:15</td>
<td>2:00</td>
</tr>
<tr>
<td>75th percentile</td>
<td>3:00</td>
<td>2:39</td>
</tr>
</tbody>
</table>

Figure 4.1: Study 1: complete data set. Plot of ATLG scores vs. time taken to complete the survey. Notice the concentration of subjects who scored 17 on the fast-side of 2:15. N = 168.
Figure 4.2: Study 1: complete data set. Plot of scores on the ATLG scale vs. overall inclination to situational attributions for the entire data set. (r = .21, p = .006, N = 168). Notice how many subjects scored 17 on the ATLG.

Figure 4.3: Study 1: How the same data looks when subjects who completed the survey in under 2:15 are excluded. (r = .53, p < .001, N = 71).
**Figure 4.4:** Study 2: complete data set. Plot of scores on the attitudes to abortion scale vs. overall inclination to situational attributions for the entire data set. \( r = .43, p < .001, N = 160 \).

![Figure 4.4](image1)

**Figure 4.5:** Study 2: How the same data looks when subjects who completed the survey in under 2:00 are excluded. \( r = .62, p < .001, N = 78 \).

![Figure 4.5](image2)
Table 4.2: Study 1: Result of excluding only those subjects who failed at least one comprehension check and were in the top quartile for survey completion speed. For this group, self-disclosure ratings and attitudes to abortion are correlated at \( r = -0.56 \) (\( p < .001 \), 95% CI [-0.69, -0.49]). \( N = 98 \). Compare to Table 2.1 on page 49.

<table>
<thead>
<tr>
<th>Item</th>
<th>( r )</th>
<th>( p )</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) ATLG</td>
<td>0.38</td>
<td>.0001</td>
<td>0.19, 0.53</td>
</tr>
<tr>
<td>CONS.</td>
<td>0.08</td>
<td>.44</td>
<td>-0.12, 0.27</td>
</tr>
<tr>
<td>(b) ATLG</td>
<td>0.10</td>
<td>.80</td>
<td>-0.10, 0.30</td>
</tr>
<tr>
<td>CONS.</td>
<td>0.05</td>
<td>.66</td>
<td>-0.15, 0.24</td>
</tr>
<tr>
<td>(c) ATLG</td>
<td>-0.31</td>
<td>.02</td>
<td>-0.48, 0.12</td>
</tr>
<tr>
<td>CONS.</td>
<td>-0.14</td>
<td>.18</td>
<td>-0.33, 0.06</td>
</tr>
<tr>
<td>(d) ATLG</td>
<td>-0.39</td>
<td>&lt;.001</td>
<td>-0.54, 0.20</td>
</tr>
<tr>
<td>CONS.</td>
<td>-0.05</td>
<td>.63</td>
<td>-0.25, 0.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Component</th>
<th>ATLG</th>
<th>( r )</th>
<th>( p )</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONS.</td>
<td>0.45</td>
<td>&lt;.001</td>
<td>0.28, 0.60</td>
<td></td>
</tr>
<tr>
<td>CONS.</td>
<td>0.12</td>
<td>.25</td>
<td>-0.08, 0.31</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: Study 2: Results of excluding only those subjects who failed at least one comprehension check and were in the top quartile for survey completion speed. For this group, self-disclosure ratings and attitudes to abortion are correlated at \( r = -0.34 \) (\( p = 0.002 \), 95% CI [-0.50, -0.16]). \( N = 104 \). Compare to Table 2.2 on page 53.

<table>
<thead>
<tr>
<th>Item</th>
<th>( r )</th>
<th>( p )</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) ATA</td>
<td>-0.45</td>
<td>&lt;.001</td>
<td>-0.59, -0.29</td>
</tr>
<tr>
<td>CONS.</td>
<td>-0.34</td>
<td>&lt;.001</td>
<td>-0.50, -0.16</td>
</tr>
<tr>
<td>(b) ATA</td>
<td>-0.32</td>
<td>.004</td>
<td>-0.48, -0.13</td>
</tr>
<tr>
<td>CONS.</td>
<td>-0.24</td>
<td>.02</td>
<td>-0.41, -0.05</td>
</tr>
<tr>
<td>(c) ATA</td>
<td>0.15</td>
<td>.17</td>
<td>-0.04, 0.34</td>
</tr>
<tr>
<td>CONS.</td>
<td>0.12</td>
<td>.22</td>
<td>-0.07, 0.31</td>
</tr>
<tr>
<td>(d) ATA</td>
<td>0.45</td>
<td>&lt;.001</td>
<td>0.29, 0.59</td>
</tr>
<tr>
<td>CONS.</td>
<td>0.34</td>
<td>&lt;.001</td>
<td>0.15, 0.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Component</th>
<th>ATA</th>
<th>( r )</th>
<th>( p )</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONS.</td>
<td>0.50</td>
<td>&lt;.001</td>
<td>0.34, 0.63</td>
<td></td>
</tr>
<tr>
<td>CONS.</td>
<td>0.37</td>
<td>.0001</td>
<td>0.19, 0.53</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.4: Study 2: Correlations for complete data set including subjects who failed speed and comprehension checks.

<table>
<thead>
<tr>
<th>Item</th>
<th>ATA</th>
<th>CONS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>-0.44</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>-0.31</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>(b)</td>
<td>-0.34</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>-0.22</td>
<td>.01</td>
</tr>
<tr>
<td>(c)</td>
<td>0.15</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>0.14</td>
<td>.10</td>
</tr>
<tr>
<td>(d)</td>
<td>0.45</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>0.28</td>
<td>.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Component</th>
<th>ATA</th>
<th>CONS.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.50</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>.0001</td>
</tr>
</tbody>
</table>

Table 4.5: Study 1: Result of excluding subjects who failed at least one comprehension check. Correlations of attitudes to gay men (ATLG scores) and conservativeness with importance ratings for items (a) - (d) and component scores for items included in factor analysis (extraction method: principal component analysis). ‘PCA loading’ indicates the component loadings (i.e., the correlations between each item and the principal component). For this subject pool, the correlation of self-disclosure ratings and attitudes to abortion scores is -0.49 (p < .001, 95% CI [-0.62, -0.42]). N = 128.

<table>
<thead>
<tr>
<th>Item</th>
<th>ATLG</th>
<th>CONS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>0.26</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>0.11</td>
<td>.20</td>
</tr>
<tr>
<td>(b)</td>
<td>0.04</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>.52</td>
</tr>
<tr>
<td>(c)</td>
<td>-0.21</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>-0.06</td>
<td>.50</td>
</tr>
<tr>
<td>(d)</td>
<td>-0.25</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>-0.04</td>
<td>.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Component</th>
<th>ATLG</th>
<th>CONS.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.30</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>.41</td>
</tr>
</tbody>
</table>