**Title**
Hijacking Reason: The Moral Ecology of Implicit Bias



**Author**
Murray, Dylan W

**Publication Date**
2017

Peer reviewed|Thesis/dissertation

**Hijacking Reason: The Moral Ecology of Implicit Bias**

by

Dylan W Murray

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Philosophy

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor R. Jay Wallace, Co-chair
Professor John Campbell, Co-chair
Associate Professor Tania Lombrozo

Summer 2017

**Hijacking Reason: The Moral Ecology of Implicit Bias**

## Abstract

Hijacking Reason: The Moral Ecology of Implicit Bias

by

Dylan W Murray

Doctor of Philosophy in Philosophy

University of California, Berkeley

Professor R. Jay Wallace, Co-chair

Professor John Campbell, Co-chair

Implicit biases operating under the radar of conscious awareness and outside the bounds of what most people would endorse seem to undermine our ordinary understanding of human agency. I argue for a particular take on why these biases' influence is so sinister—they "hijack" control of not only our actions but our very processes of practical reasoning and deliberation away from our selves and what we care about and value. I also argue that some implicit attitudes themselves count as values and forms of caring, such that our selves are partly implicit, or unconscious. This take on implicit attitudes is informed by how they work within particular social and geographic environments—that is, by their "moral ecology." For example, implicit attitudes are central components in the mechanisms that sustain cycles of endemic poverty in U.S. inner cities. Somewhat counterintuitively, I conclude that the best way to combat the hijacking of reason is often precisely more hijacking, guided by research on the nature of implicit bias and the moral ecology of the environments in which it operates.

# Contents

## Acknowledgments

For gracious discussion of and feedback on the ideas in this dissertation, I'm extremely grateful to R. Jay Wallace, John Campbell, Tania Lombrozo, Lauren Olin, Dan Khokhar, Eddy Nahmias, Manuel Vargas, Alex Madva, Véronique Munoz-Dardé, David Harding, Joshua Knobe, Antonia Peacocke, Nick French, Eugene Chislenko, Raymond Banks and the members of our Social and Political Philosophy of the American Inner City class, Quinn Gibson, Alex Kerr, Sara Gottlieb, Jonathan Phillips, Joshua Greene, John Doris, Ryan Murray, Greg Murray, Kathy Winnett-Murray, and audiences at the University of Pennsylvania Wharton School, the California Institute of Technology, and at the University of California, Berkeley.

# Chapter 1

# Introduction: Moral Ecology

We have the sense that most of our actions are self-governed, or *autonomous*—that they are "up to us" rather than others or external forces. This sense is familiar enough from ordinary, everyday life, but explaining just what autonomy (or freedom, or responsibility) *is* proves difficult.[1] One approach is to focus on what undermines or threatens autonomy; we often understand best how something works by seeing how it breaks.

Philosophers have focused on the question of whether causal determinism (the thesis that, necessarily, given the past and laws of nature, everything that actually happens has to happen) would undermine autonomy. Incompatibilists hold that *no* actions can be autonomous if determinism is true. Compatibilists claim otherwise. Following Strawson (1962), many claim that whether or not determinism is true has no bearing on the type of autonomy operative in, and important to, ordinary social life. Imagining ourselves as jurors, perhaps: even if the most famous of contemporary physicists were to deliver expert testimony to the effect that our universe is deterministic, that shouldn't affect our verdict about someone's autonomy with respect to any particular action one jot. It would not amount—in any and every case—to an automatic "not autonomous" verdict. Empirical work in experimental philosophy arguably confirms that the type of autonomy that ordinary people see as relevant—ordinary users of the concept—does not make any assumptions about (in)determinism.[2]

This classic debate between compatibilists and incompatibilists continues to smolder, and always will, but recent years have also seen increased interest in other ways that autonomy might break. The threats outside physics that have received the most attention from philosophers come from neuroscience. Benjamin Libet (1985) showed that EEG (electroencephalography) readings can predict when people will make a decision a little less than half a second before they're aware of making it, and John Dylan-Haynes and collaborators have shown that fMRI (functional magnetic resonance imaging) readings can predict with 60% accuracy whether a person will decide to push a button on the left or a button on the right

[1] I take autonomy, free will, and moral responsibility to be closely related, and will tend to leave the differences between these notions in the background until they become relevant.

[2] See Murray & Nahmias (2014), Nahmias and Murray (2010), and Nahmias et al. (2005, 2006, 2007). For dissenting evidence, see especially Nichols & Knobe (2007) and Nichols (2011).

*7-10 seconds* before the person is aware of deciding which button to press (Soon et al. 2008). Dan Wegner (2002, 2008) and John Bargh (2008), among others, have argued that these and other neuroscientific results suggest that our behaviors are caused by different neural and psychological states than those we associate with autonomy. In experiments like Libet's and Haynes', it seems that one's unconscious mental states are *bypassing* participants' conscious mental states—in particular those associated with autonomy, like their practical reasoning, deliberation, and what people value or care about (Murray & Nahmias 2014). If one's unconscious mental states systematically cut one's conscious self out of the causal chain that leads to one's own behavior in this way, autonomy seems entirely illusory.[3]

Like determinism and other threats from physics, though, the neuroscientific threat—if real—would constitute another global threat to autonomy. That is, if it has anything to say about any particular verdict, it says the same thing across the board: "not autonomous." Even if some form of global skepticism is true, though, there's still reason to investigate more fine-grained ways that autonomy might break—factors that might make us non-autonomous (or less autonomous) in some cases, but not others. That type of local threat promises to carve closer to autonomy's joints, giving us a better understanding of how it works, than global threats. Not only does focusing on local threats give us a closer rendering of autonomy's nature, but these threats should also be of more interest to moral philosophers working in practical ethics and political philosophy. It's these more contingent threats that we might actually be able to do something about or alleviate. (What tangible good is there, after all, in knowing how something breaks if that doesn't teach us how to put it back together?) In particular, taking a line broadly from Strawson (1962), Wallace (1994), and others, I'll suggest that we can learn a great (applicable) deal from thinking about how interpersonal interactions and geographic context might impact autonomy: that is, how autonomy is affected by one's particular "place" in the world. Philosophers have discussed these topics to some extent, but not as I aim to here.

John Doris (2002), Gilbert Harman (1999), and other moral psychologists have discussed work from situationist social psychology showing that seemingly normatively irrelevant features of situations—like whether or not one just found a dime—can affect moral behavior, like whether one offers to help someone pick up a dropped stack of papers shortly thereafter. To many, these results have seemed to challenge our ordinary sense of ourselves as autonomous agents. If our actions are frequently influenced so much by so many tiny features of the situations we find ourselves in, what contribution is left for us to make, as individual agents who move from situation to situation?[4] Situationism, though, primarily

[3]For criticism of these neuroscience-based arguments, see Mele (2009, 2013) and Nahmias (2014). It's possible, for instance, that participants in the Libet and Haynes experiments are simply waiting to see which option they "feel" (more) like doing (or when), that the neural imaging reveals this emerging feeling (not a decision), and that people then make their decision on the basis of the feeling at the time they're aware of doing so (see Trevena and Miller 2009 and Pockett and Purdy 2010). In that case, the findings wouldn't challenge our sense of autonomy at all: making our decisions on the basis of what we feel like doing (where that feeling itself is simply given, and not chosen), is perfectly familiar and unthreatening.

[4]See also Nahmias (2007), Nelkin (2005), Vargas (2013b), Murray (2015), and Murray (forthcoming).

focuses on very *small, surprising* features of situations—dimes, the presence of bystanders, and the like. I aim focus here on the *large, obvious* features of social situations and place instead—especially those of most interest to moral and political philosophers with pressing practical aims.[5] One all-too-common feature of many social situations is manipulation.

Even in the classic free will debate, the most prominent recent argument for incompatibilism alleges that determinism's effect on autonomy just is the same as being intentionally manipulated by another agent—e.g., through direct hypnosis or brainwashing. In these "manipulation arguments," we're asked to compare the effect of determinism to social indoctrination or conditioning (Kane 1996, Pereboom 2001, 2014). And other philosophers have addressed the question of whether, in principle, culture and social circumstances—like living in conditions of extreme poverty—might affect autonomy (Moody-Adams 1994, Walker 1969, Wolf 1987). Setting determinism aside, though, there's still an independent question about whether, and if so how, social and geographic circumstances can be genuinely manipulative in the first place (in the way that being brainwashed or hypnotized to perform a particular action by another agent or determinate group of agents can be). A first step in tackling this question is figuring out just which features of manipulation undermine autonomy, which some experimental philosophers have started to address.[6] This research suggests that manipulation of the hypnosis and brainwashing variety intuitively undermines autonomy because a manipulator's intentions are perceived to interfere with the causal connection between manipulees' own intentions and manipulees' actions. Especially threatening are cases where manipulators bypass a manipulee's own decision-making processes in the causal chain that leads to and controls the manipulee's action.

Drawing on interventionist theories of causation,[7] Murray & Lombrozo (2017) suggest the results above show that manipulation is threatening because, and to the extent, that it involves a decrease in the *counterfactual dependence* of the outcome on the manipulee's relevant mental states, a working assumption I adopt here. Bypassing is the limiting case, in which the dependence of the outcome on the manipulee's mental states is reduced to zero. But other types of interpersonal influence can decrease counterfactual dependence to varying lesser extents, explaining why they mitigate autonomy to varying lesser degrees.[8]

The question then is whether more mundane, everyday varieties of "manipulation" of the sort commonly seen in the real world can decrease the dependence of people's actions on their mental states in the same way involved in canonical, "science-fictional" cases of manipulation, if to a lesser degree. I'll argue that a large body of scientific work shows that

[5]This difference in emphasis shouldn't be over-sold, however. Other classic situationist findings like the obedience experiments of Milgram (1974) and Zimbardo (2007) do feature large, obvious features of situations. Indeed, these experiments (inadvertently) study the effects of manipulation.

[6]See Woolfolk, Doris, and Darley (2006), Sripada (2012), Feltz (2012), and Phillips & Shaw (2014).

[7]See, e.g., Spirtes, Glymour, and Scheines (1993), Pearl (2000), Hitchcock (2001), Woodward (2003), Woodward & Hitchcock (2003), Hitchcock & Woodward (2003), and Campbell (2010).

[8]Intuitively, the more counterfactually dependent an action or outcome is on the manipulator's mental states, the less counterfactually dependent it is on the manipulee's. I return to this research in more detail in Ch. 2. Note also that Murray & Lombrozo (2017) and the papers referenced in notes 2 and 6 don't ask experimental participants directly about 'autonomy' but instead about its various cognates (see n. 1).

they can. The foundation of this argument is research showing that our own unconscious can "manipulate" us, or at least reduce counterfactual dependence on the mental states that moral autonomy properly depends on. We can then extend this picture by seeing how social and geographic—ecological—influences can co-opt these unconscious mechanisms.

The behavioral and social sciences are starting to amass extensive evidence of just this type of unconscious influence: widespread real-world interference in our thought and behavior that resembles manipulation. At its psychological base, the general *dual-processing* framework that this evidence comes from is now nearly ubiquitous within cognitive science, and I adopt it here—in its broadest outlines (Kahneman 2011, Greene 2013). According to dual-processing theories, psychological processes and attitudes fall into two categories: System 1 *implicit attitudes*, which tend to be unconscious and associative and influence thought and behavior "behind the scenes" or "under the radar"; and System 2 *explicit attitudes*, which are the typically conscious, often rational or reason-based thoughts that our waking lives seem, "from the inside," to be governed by. The dual-processing research, however, purports to show that our behavior is often driven by our implicit, rather than explicit, thoughts and attitudes, which many researchers take to threaten autonomy.

It's now been demonstrated that a whole host of everyday behaviors are shot through with the influence of implicit biases and other attitudes that radically conflict with what we explicitly—at the System 2 level—think we ought to do, or think there's even any reason to do (Greenwald and Banaji 1995, Banaji and Greenwald 2013). Of course, there are other obstacles to autonomy in this vicinity that philosophers have discussed: addiction, weakness of will, temptation, sour grapes, compulsion, delusion and other mental illnesses, and, most relatedly: rationalization, self-deception, and motivated reasoning. The dual-processing results, however, seem to reveal a more sinister type of influence: that we're often manipulated by our own implicit attitudes, or at least affected in a way that approximates intentional manipulation by another agent. While other philosophers have noted the similarity (Doris 2015, H. Smith 2015), here I try to take the analogy to manipulation as far as it can go.

Philosophical attention has focused on implicit attitudes' ability to influence action through separate causal routes that bypass explicit thought entirely. But I argue that the more sinister cases occur when implicit attitudes influence action *through* explicit deliberation and reasoning. For example, most people recommend college admission more to white than black applicants, but they also tend to rank as more important for admission whatever qualifications the white applicants happen to have. If a white applicant has high grades but a black applicant has high SAT scores, then most people rank grades as more important than SAT scores (for admission in general); if the reverse, then vice-versa. We typically wouldn't expect a brain tumor or anything wholly inanimate to "know" how to manipulate or coerce one's deliberation in these quasi-intentional, semantically-sensitive ways. In contrast, the neuroscientific results and other obstacles to autonomy like addiction and temptation don't testify to the existence of these sorts of quasi-goal-driven behavior on the part of the unconscious. By analogy: to stop, sabotage, or otherwise get around or ground an airplane is one thing, but to actually use the plane for your own purposes—to hijack it—you'd better *know how to fly it*. There must be some sense in which implicit attitudes are "intelligent" enough

to maneuver one's explicit attitudes, in a way that succumbing to a brain tumor, weakness of will, or bypassing of one's explicit attitudes entirely doesn't require.

The counterfactual dependence framework can be used to spell out (degrees) of manipulation in terms that are not inherently intentional or teleological, as I argue at the end of Chapter 2. As such, the framework can naturally be extended to cover the sense in which (degrees of) hijacking by our own unconscious implicit attitudes undermines autonomy (to various degrees). Chapter 3 delves into just how deep hijacking goes and how it plays out over time, drawing on allied work in behavioral economics to show that one's very conception of what things are good (valuable) and which things are bad (disvaluable) can even be hijacked. Later chapters go on to discuss how social and geographic context can induce and co-opt the hijacking mechanism, thereby affecting autonomy in a way that's mediated by the implicit "manipulative" psychological processes detailed in earlier chapters.

In Chapter 4, I argue that implicit attitudes sometimes enhance rather than diminish autonomy. Overlooked research shows that a unique subclass of implicit attitudes involve a capacity to care about and value other things by associating *oneself* with those things: these attitudes are ways of making oneself emotionally invested in and vulnerable to such things' well-being on the implicit level. Because these particular implicit attitudes are subject to pressures to form a more coherent outlook than one's other implicit attitudes (one should not bind one's implicit well-being up too closely with both the Hatfields *and* the McCoys), these implicit forms of valuing are partly constitutive of the attitudes in virtue of which one can be autonomous, rather than interfering with the effect of such attitudes on behavior.[9] Indeed, when it's these implicit forms of valuing that hijack one's behavior away from one's explicit values, one's considerable less "off the hook" than when hijacking is driven by other implicit attitudes (that don't constitute implicit forms of valuing).

While the dual-processing research thus provides some propitious news about autonomy, it's not just the findings from cognitive science and behavioral economics that raise the hijacking threat. I suggest in the closing chapters that the dual-processing results are just the tip of the iceberg once we understand implicit attitudes within the ecological contexts in which they operate. Typically, humans "contract" whatever implicit attitudes and biases we end up having from others in our social environment—often through the non-intentional expression of *their* implicit attitudes (Mackie et al. 1996). Indeed, implicit attitudes seem to be a fundamental and unavoidable feature of humans' whole basic scheme of drawing generalizations and tracking the probabilistic relationships between different aspects of the environments in which we live (Leslie forthcoming and Gendler 2011).

An entire self-perpetuating cycle operates outside of anyone's conscious intent or even awareness, let alone endorsement—as if our own implicit attitudes are "trying" to get their way with us. For one, implicit biases can hijack *others'* reasoning, not just one's own, and in social concert, they can create and sustain relationships of serious exploitation—e.g., as some of the mechanisms that drive continued racial discrimination in the US housing market. The foregoing chapters partly explain how social forces can have these sorts of coercive

[9]These attitudes thus count as *caring* in the sense articulated by Frankfurt (1999) and Bratman (2007).

and manipulative effects even when they're wholly unintentional. As my main case study, I'll argue in Chapter 5 that implicit attitudes constitute an integral part of the vicious cycles endemic to much enduring poverty—specifically, the type of intergenerational "poverty trap" in which many urban African Americans still find themselves. Implicit biases—not just racial—segregate the urban poor into *ghettos*—areas of particularly concentrated poverty and associated forms of disadvantage. This geographically and socially concentrated disadvantage then comes to be associated with blacks themselves, such that implicit biases against place (and class) combine with those based on race to further exacerbate discrimination-based residential segregation. This increased bias ratchets up the cycle, leading to increased concentrations of poverty. Moreover, many of the negative effects of living in poverty on the poor that perpetuate the cycle—like over-borrowing—are mediated by implicit attitudes and mechanisms (Mullainathan and Shafir 2013). In closing, I'll suggest how we might think about alleviating some of these effects, specifically with an eye to intervening on the implicit level, and whether it might be justified to intentionally use the dual-processing results to hijack or manipulate ourselves as a matter of public policy.

Ultimately, I mean to vindicate the claim that there are more interesting topics for philosophers working on autonomy to turn their attention toward than causal determinism. Expanding on Nahmias (2007), I hope to show that we can learn more about the contours of autonomy—about just what it is and how it works, as well as how much of it we've got—from the social sciences than we ever could from (meta)physics. More fundamentally, I aim for this to constitute the first foray into what might be called *moral ecology*.

*Moral psychology* in its recent, cross-disciplinary sense has seen considerable success (e.g., Greene and Haidt 2002, Doris et al. 2010). But practitioners have come to see the limits of what can be accomplished in the laboratory. Morality, especially autonomy and the related practices of holding one another responsible, are inherently social phenomena. One can ask people about morally loaded scenarios that involve multiple actors, and even bring multiple participants into the lab at once, but it's difficult to find room for entire societies or peoples. Thankfully, we can extend the reach of moral psychology without backsliding on its scientific credentials. Sociology and the other behavioral sciences lie ready in wait.

If moral psychology is "the study of morality in its psychological dimensions" (Wallace 2007), moral ecology is the study of morality in its ecological dimensions—in particular, those concerning moral interactions between agents within (and across) environments. Within biology, ecology is the study of organisms and their interactions within particular environments, but with an emphasis on drawing generalizations across and providing explanations of differences between such environments—e.g., differences between food chains, or the common elements of the water cycle. Correspondingly, moral ecology is the study of moral agents and their interactions within environments with an emphasis on drawing generalizations across environments and explaining differences in interactional patterns between them.

Just as (biological) ecologists often focus on phenomena that simply cannot be understood without understanding how organismal interactions are spatiotemporally located (both in the sense of particular location and the sheer fact of having any such location at all), so moral ecology focuses on phenomena of moral import that cannot be understood indepen-

dent of spatiotemporal, geographic context. I'll argue that the cycle of poverty characteristic of U.S. ghettos—including the associated moral interactions surrounding them—is precisely such a phenomenon (Duneier 2016, Shelby 2016). To understand the continued persistence of black ghettos—despite sustained drops in levels of explicit racism since the 1960s—requires understanding the role of *implicit* bias in continued residential segregation. But we cannot understand how these same biases are in turn exacerbated, ratcheting up residential segregation, without understanding how they lead to the geographic *concentration* of poverty and its associated disadvantages, including the behaviors these tend to give rise to. Concentration is an inherently ecological phenomenon, and it constitutes another crucial part of the cycle of poverty. This has been one of the primary theses of urban sociology (Massey & Denton 1993, Sampson 2012). This is but one particularly large, obvious example of how, without both the cognitive science and the sociology, we cannot as moral philosophers make sense of some of the most pressing real-world moral concerns of our time. That is, for an applicable moral philosophy in this and many other domains, we need to understand how implicit biases and other psychological attitudes operate in the wild. We need a moral ecology.

To understand autonomy in the way required to address most questions of public policy (e.g., regarding poverty and the housing market) requires looking to the entire range of the behavioral and social sciences, not just psychology, and it requires looking beyond small, seemingly insignificant features of situations to larger, more obvious ones. Many of the results documented in what follows are relevant to and should be taken into account in such decisions, though this should be done with caution. In the closing chapter, I'll discuss questions of policy, with the specific focus of how we might intervene on people's implicit attitudes in ways that affect autonomy. I'll be especially concerned with *nudges*, Richard Thaler and Cass Sunstein's (2008) proposal to use dual-processing research to shape "choice architecture"—setting people's option sets up so as to promote socially desirable choices. For instance, changing driver's licenses to require checking the box on the back to opt out of organ donation, rather than in, dramatically increases the number of donors merely by changing the default, which System 1 implicit attitudes are biased in favor of.[10]

The main worry so far leveled against nudges is that they undermine autonomy—some opponents even going so far as to suggest that nudges amount to governmental mind-control and manipulation of its citizens. The preceding chapters suggest this is misguided: we're often manipulated or hijacked anyway. And many of Thaler and Sunstein's nudges are intended to dislodge just such extant implicit interference, or naturally occurring "nudges" (though these, of course, also include those introduced to human environments by modern marketing and a whole host of other intentional and unintentional acts). The question then isn't so much whether to nudge or not as which way, and who gets to decide. Should we use planned governmental intervention, subject to the safeguards of electoral representation, or should we leave the naturally occurring nudges already there undisturbed?

In many cases, nudging should be uncontroversial. We should try to assist the poor in overcoming their own implicit attitudes that keep them in poverty (like those responsible

[10]See, e.g., Kurtz & Saks (1996) and Johnson & Goldstein (2003).

for over-borrowing), and to remove or mitigate implicit biases based on race, socioeconomic class, and neighborhood lived in. The situation is more complex, however, when we consider nudges that might affect what people implicitly care about—what they've bound up with their own implicit self-worth, and so who they are on an implicit level. Interventions that affect one's identity in these ways are subject to more presumptive moral objections (Appiah 2005, Shelby 2016). I take this to apply on the implicit level, as well. Attempts to nudge people in ways that change what they implicitly value, and so their implicit identities, raise deeper objections, and at least need to meet a higher justificational bar than nudges that affect other implicit attitudes. Most such nudges are (as yet) purely hypothetical, but one real-world example: some authors have suggested changing the default neighborhoods that section 8 and other housing voucher recipients move to.[11] Nudges like these may change the implicit identities of those the policy affects—in this case, by "nudging" people literally out of certain neighborhoods and ways of life. At the very least, any attempt to press dual-processing research into serious public policy service in many of the most pressing situations requires an appreciation of the situational, ecological factors in play.

Before delving into the cognitive science of implicit attitudes—the foundation for the moral ecological framework to follow in later chapters—a note on style: The vexations raised by implicit attitudes, in the lab and out, are fundamental to our everyday lives and interactions with each other; indeed, I think they force us to look at ourselves and one another somewhat differently, and not all for the best. At the same time, many of the empirical details—at all levels of magnification—no doubt remain to be filled in. This raises questions of scale, and of how much precision to ask for (a broadly Aristotelian theme). It'd be a shame not to write both accessibly and rigorously when writing about a topic like this—and I think philosophy is particularly well-placed to explain the basic worries these results raise in terms of a sense of agency that people recognize themselves as holding, at a level of generality into which further empirical findings can be slotted. I want something between a nature documentary ('Autonomy in the Wild', or 'Implicit Attitudes Attack!') and the pocket 'Field Guide to Implicit Biases'—something that traces the natural arc of the empirical research while also providing some general taxonomy of the relevant threats to (and species of) autonomy.[12] Our story starts in the psychology lab, but we'll cross campus to the economics department and school of management, then have urban sociology guide us through the inner city. We'll close with the policy wonks on the steps of government, but refrain—as more documentaries might—from making any (controversial) normative recommendations. The hope is that a mature moral ecology might allow others to go on to do so.

[11] See, e.g., de Souza Briggs, Comey, and Weismann (2010) and de Souza Briggs (2008).

[12] Such a field guide would have its spine number alongside entries on all the more familiar threats to autonomy traditionally discussed by philosophers, like temptation, weakness of will, and addiction.

# Chapter 2

# The Hijacking of Reason

## 2.1 Introduction

Imagine you're on a college admissions committee, choosing which candidates to recommend. You look over each application, assessing its merits compared to the others you've seen, and consider its strengths and weaknesses. Then you put it in the "Recommend for Admission" stack or the "Don't Recommend" stack. Occasionally you rethink a decision and move one of the applications from one stack to the other. But all the while, you have every sense that your decisions about whether to recommend candidates for admission or not are under your own control: they're based on *your* reasoning and assessment of the merits of each applicant. That is, each action feels like it's governed by you rather than others or external forces—it feels like an action you're autonomous with respect to.

But now, suppose it turns out that you were hypnotized several days ago by evil neuroscientists bent on influencing this very admissions process. They covered their tracks just as expertly as they designed your new mental states, and you have no indication or memory of their influence. Specifically, imagine these neuroscientists tampered with your reasoning and assessment of each candidate in a way that led you to prefer white applicants when compared to black applicants. It's not that they instilled any new explicitly racist thoughts in you, but that they affected the relative importance that you placed on different dimensions of applications such that you'd end up recommending more white applicants than black.

This type of intentional manipulation by another individual or set of individuals is a quintessential threat to autonomy, albeit thankfully science-fictional. In the imagined scenario, you have every sense that your admissions recommendations are up to you, and express *your* views about what's relevant and important for college admission. But this is illusory.

Compare this to a second scenario, otherwise identical to the first except that instead of another agent manipulating you, your own unconscious mental states bias you in favor of white and against black applicants. You have no idea these attitudes are there. Just as in the first scenario, you take yourself to be sincerely egalitarian. Nonetheless, these attitudes influence your assessment of the candidates' criteria in making your admissions

recommendations in a way that favors white applicants. This is a real scientific finding.

Hodson, Dovidio, and Gaertner (2002) asked participants to make college admission recommendations for candidates who had either strong high school grades and SAT scores, only strong grades, only high SAT scores, or neither, and who were either white or black. For the strongest and weakest candidates, implicit biases had no significant effect on admissions decisions. But among those with mixed credentials, white candidates were recommended for admission more than black applicants. Moreover, these participants reported that they took whichever credential the black candidate was weaker in to be more important for college admission. That is, when a black candidate had only high SAT scores, participants took strong grades to be more important for admission (in general), but when the black candidate had strong grades, participants instead took high SAT scores to be more important for admission. It's as if one's own unconscious is making one reason in a particular way.[1]

These results are driven by implicit biases, a notion I go on to develop below. In this chapter, I argue that such biases and other implicit attitudes can undermine autonomy in much the same way that intentional manipulation by another individual agent can. First, some preliminary remarks about the notion of autonomy are in order.

There are different ways of understanding autonomy or "self-governance," of course, and these differences will become relevant. To set things up, though, I'll simply adopt a particular approach to autonomous agency: the *self-expressive* approach. According to this position, we're autonomous with respect to actions to the extent that they express our selves.[2] The psychological attitudes that are most constitutive of the self are those that comprise what one cares about, values, or "identifies" with.[3] I mean to stay relatively neutral between different versions of the self-expressive theory for now, but Bratman's is among the most well-developed and so gives some indication of such theories' resources.

According to Bratman (2007), we're autonomous when we're guided by certain plans or intentions, where these are understood to be a type of valuing. Specifically, these are plans to treat one's first-order desires as reason- or justification-providing in one's (motivationally effective) practical deliberation—treating them as having different weights or importance in one's deliberation about what to do. In other words, valuing something is not simply a matter of desiring it, but of desiring it in a way that one would use to justify one's action if called upon to do so.[4] Finally, in order for *behavior* to be autonomous, it must not only conform to one's valuing, it must (non-deviantly) *express* it, which at a minimum involves being caused by the valuing rather than some other source (Bratman 2007: 70). According to Bratman, these valuings constitute the agent's unified perspective on practical matters,

[1]The Hodson et al. (2002) results compare decisions and ranking of qualifications across different conditions of the black applicant. See also Dovidio and Gaertner (2000), Ullman and Cohen (2005), and Son Hing et al. (2008). These results are discussed further in Section 3.

[2]I assume autonomy comes in degrees—see, e.g., Arpaly (2003), Arpaly & Schroeder (2014), Capes (2013), Coates & Swenson (2013), Khoury (2014), Nelkin (forthcoming), and Tierney (2013).

[3]On identification, see Frankfurt (1971) and Watson (1975). Other self-expressive approaches include Aristotle (2004), Hume (1738/1978), Dewey (1957), Frankfurt (2006b), Bratman (2007), and Sripada (2016).

[4]Valuing must also be an intrinsic desire and non-fungible.

and in this sense (have the authority to) "speak for" the agent's self.

I return to self-expressive theories in Chapter 4, but this is enough to see why implicit biases might threaten autonomy in a way akin to intentional manipulation by another agent. If what's threatening about manipulation is that it takes control of our actions away from our own *selves* (e.g., our valuings), then it seems that at least some implicit attitudes can have the same effect on our control over our actions. Moreover, in a way that other forms of practical shortcoming simply don't provide evidence for, it's not just people's overt behaviors that are affected by their implicit attitudes, but in some cases their very practical reasoning.

John Doris (2015) has perhaps come closest to advancing the view I'll argue for here. Doris understands morally responsible agency in terms of autonomous agency, as I'll go on to in Chapter 4, and Doris (2015: 28, n. 5) adopts a more congenial version of a self-expressive approach for talking about the dual-processing findings—borrowing from Bratman but not requiring valuings or cares to be explicit or reflective—a qualification I also hereby adopt. However, Doris argues only that these results undermine theories that take reflective deliberation to be required for autonomy, whereas I think Hodson and colleagues' and similar research cuts deeper, in a more sinister way. It's not just that participants do not *reflectively* take the effects of their implicit biases on their ranking of admissions criteria to be rational, or justified, but that they don't take these to be reasons at all. It's not just that our implicit attitudes sometimes lead us to act in ways that we don't reflectively, deliberatively take to be justified all-things-considered, but that they often lead us to act in ways that we don't take to be reasonable or justified at all (on any level, reflectively or otherwise).

Most discussions of the dual-processing findings have focused on cases in which implicit attitudes conflict with and "beat" explicit, reflective attitudes to produce behavior. Joshua Greene (2001, 2013) has discussed ways in which one's gut affective reactions (e.g., aversion to muscular force) can influence moral judgments, and Tamar Gendler (2008a, 2008b) has discussed how implicit associations can lead, e.g., to a fear response when suspended on a transparent platform above the Grand Canyon, even though one explicitly believes that one is completely safe and has no reasons to be afraid. Doris focuses on cases in which implicit, System 1 attitudes influence behavior in a way that "bypasses" one's explicit, System 2 attitudes entirely. I think Doris is right that bypassing comes closer to characterizing what's uniquely unsettling about the dual-processing results, compared to cases of simple comeuppance or "beating." But I think there's a further type of interaction between implicit and explicit attitudes that's even creepier, and that so far most researchers have overlooked.[5]

Sometimes, implicit biases "get their way" with us not by beating or bypassing our explicit, deliberative thought, but by *hijacking* it. In the Hodson et al. (2002) experiments, it's not that one doesn't reason about the applicants at all, nor that the conclusion of one's reasoning is overpowered by one's implicit bias. Participants' very evaluative judgments and reasoning about which criteria are comparatively more important for college admission (in general—not just about a particular pair of candidates) is being "manipulated" by their implicit biases. Participants' explicit reasoning precisely isn't overcome or bypassed

[5]Though Doris (2015: 64) comes close—e.g., asking "What other non-reasons might infect deliberation?"

in producing behavior in these cases; instead, their reasoning itself has become a mere rationalization of whichever "goal" their implicit bias has. One's implicit bias seems to have replaced one's self in the driver's seat of one's own practical reasoning. It's in such cases of hijacking, I believe, where the analogy with manipulation is strongest.

After introducing implicit attitudes and dual-processing theories in more detail in the next section, I argue that hijacking should be added to our catalogue of the ways in which Systems 1 and 2 can interact in Sections 3 and 4. In the limiting case—such as the Hodson et al. (2002) study—the *only* practical reasoning one engages in, and so the only considerations one might even potentially proffer as justifications for one's choices and actions—are implicitly-produced rationalizations, rather than responses to reasons or anything one values. In Section 5, I make good on kicking away the analogy to intentional manipulation by explaining how both manipulation and hijacking by one's implicit attitudes threaten autonomy by sharing in (degrees of) literally the same underlying property, which I cash out in terms of interventionist theories of causation. Manipulation is threatening because it involves intentional interference, but *intentional* interference is primarily threatening because it (typically) undermines the counterfactual dependence of a manipulee's action on what that manipulee cares about and values to a large degree. Implicit attitudes exercise an intermediate degree of interference with this counterfactual dependence in cases of hijacking, explaining their intermediate, "quasi-manipulative" impact on agents' autonomy.

## 2.2 Dual-Processing: Implicit Bias

There are in fact several semi-separate, individually massive literatures on dual-processing. Most of these trace back to Amos Tversky and Daniel Kahneman's pioneering work on "heuristics and biases," which demonstrated numerous instances in which our use of mental shortcuts or rules of thumb predictably leads to a whole stable of systematic errors.[6] System 1 implicit attitudes offer up quick and dirty answers to problems, and are always operating. Often, System 1's answers are normatively correct. System 2 explicit attitudes only come online to check these answers sometimes, and these attitudes and processes are slower and require more effort to engage in. In the increasingly canonical example, consider the first answer that pops into your head when asked: "A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?" If you're like most people, you have an initial reaction of "10 cents" (your System 1 answer) which is easy to see is incorrect once you step back and engage your reflective, explicit System 2 attitudes. You can use your explicit attitudes to correct the answer once you notice your initial mistake, but in many cases, you never do. It's widely believed that to accommodate these and many other heuristics and biases findings, in which answers or other behaviors systematically diverge from the normative standard, we need to posit two different mental systems (though

[6]The original work is summarized in Tversky and Kahneman (1974, 1981); I discuss the classic heuristics and biases in more detail in the next chapter. For a more complete history of the dual-processing literature, including dual-processing theories of memory, see Evans & Frankish (2009).

this need not mean more than positing two different types of mental process). The work that Tversky and Kahneman's research has inspired has been somewhat segregated between psychology and behavioral economics. At the risk of perpetuating artificial boundaries, I primarily discuss the former research in this chapter, and the latter in the next.[7]

Dual-processing theories sort cognitive processes into two categories: System 1's *implicit attitudes*, which are typically unconscious, effortless, automatic, evolutionarily older, and only capable of integrating information about one or a limited number of an object's attributes; and System 2's *explicit attitudes*, which are typically conscious, effortful, controlled, evolutionarily more recent, and capable of comparing objects to each other along several attributes (by weighing these attributes against one another). System 2 comprises the thought processes that we're typically consciously aware of, and which we tend to think of as delivering the normative, rational answer. In contrast, System 1 is less visible to the naked eye, operating under the radar, and was initially posited precisely to explain the relevant findings: it's the source of the heuristics that derail rational thinking and lead to biases, like the "10 cents" answer. (Again, though, while System 1 may be subject to systematic mistakes, it too typically delivers normatively correct responses.)[8]

There's debate about which pair of features, if any, is definitive of the distinction, but many theories mark it in terms of different types of mental transition: implicit System 1 attitudes function *associatively*, explicit System 2 attitudes operate according to rational rules.[9] Nomy Arpaly (2005, 2006: Ch. 2) distinguishes three types of causal transition: *brute physical causation* (e.g. thinking 'Q' because you're hit on the head), *(merely) content-efficacious causation* (e.g., thinking 'Q' after hearing 'P' because they've been "constantly conjoined" in the past and you've come to associate the two), and *reasons-responsive* or *rational causation* (e.g., inferring 'Q' as the result of believing 'If P, then Q' and 'P'). The last two are instances of mental causation—occurring in virtue of the meaning or content of the mental states involved—but only the latter occurs in virtue of the logical or other normative relations between those contents. Associative relations are perhaps the primary example of how there can be content-efficacious, but not reasons-responsive, causation.

So understood, implicit attitudes are not mental states with propositional contents—'black is bad' or 'white is good'—and they need not involve evaluative or normative concepts. Instead, *implicit attitudes* (including *implicit biases*) are associative relations of varying

[7] I have other reasons for this separation as well: the social psychological findings are easier to feel the intuitive "manipulative" pull from, whereas the findings from behavioral economics allow one to see just how deep the threat goes—showing that implicit attitudes can even infect one's current conception of what's good in the first place (and bad), thus shaping behavior and cognition over temporally extended periods.

[8] For other comprehensive discussions of the divide, see: Baron (2007), Chaiken and Trope (1999), Evans (2007), Evans and Over (1996), Gilovich, Griffin, and Kahneman (2002), Kahneman, Slovic, and Tversky (1982), Sloman (1996), Stanovich (2004, 2010), and Wilson (2002). Note also that implicit "attitudes" need not be stable or robust across situations in the way that (explicit) attitudes typically are. For this reason, a more appropriate label might be the simpler 'psychological state', but I'll follow standard terminology here.

[9] See, for example, Fazio (2007), Gawronski and Bodenhausen (2011), Gendler (2008), and Rydell and McConell (2006). Kahneman (2011: 13) seems to stress the effortful/effortless contrast when distinguishing the two systems, but he identifies "associative memory" as "the core of System 1."

strengths between one concept and another concept, or between one concept and an affective valence, whereby the activation of one increases (or decreases) the likelihood of the other's activation (e.g., the way SALT tends to activate PEPPER, and vice-versa).[10]

Numerous sources of evidence demonstrate implicit attitudes in this associative sense:

- *Semantic Priming*: Bargh et al. (1996) argue that priming participants with the concept ELDERLY causes them to walk more slowly, and Williams and Bargh (2008) argue that holding warmer drinks (activating WARM) causes participants to be both more friendly to others and to rate them more favorably on a subsequent questionnaire.

- *Sympathetic Magic*: People tend to disprefer beverages into which completely sterilized cockroaches have been dipped (Rozin et al. 1986). They'll also pay more for sweaters if they were worn by celebrities (Nemeroff and Rozin 1994), but less so if they've been washed (Newman et al. 2011), all ostensibly because of associative "contagion."

- *Name-Letter/Implicit Egotism Effects*: Virginia and Georgia are 36% more likely to move to states with those names than others; Geoffrey and George 42% more likely to become geoscientists due to positive associations with their names (Pelham et al. 2002). More generally, people tend to implicitly prefer things associated with themselves.[11]

- *Implicit Association Test (IAT)*: A vast number of studies using the IAT have demonstrated the existence of widespread implicit racist, sexist, and various other prejudicial biases that often conflict with explicit (self-reported) attitudes (e.g., Greenwald and Banaji 1995, Banaji and Greenwald 2013). The majority of white Americans, e.g., find it easier to quickly categorize positive words (and pictures) with white faces and negative words (and pictures) with black faces than *vice versa*, even when they explicitly (and probably honestly) disavow any racist attitudes. These results are widely thought to be due to participants' having stronger implicit associations between BLACK/BAD and WHITE/GOOD than between BLACK/GOOD and WHITE/BAD.

Because most Americans associate more negative affect with BLACK than WHITE, they're quicker to categorize the former with other concepts also associated with negative affect than the latter. Implicit affect can in turn affect overt behavior: agents are automatically attracted or repelled from courses of action to varying degrees, depending on the strength

[10] I use all-caps for names of concepts throughout (e.g., SALT and PEPPER). There's further debate about whether the mental *states* or attitudes so related by these types of causal transition have internal associative structure. Some, like Mandelbaum (2013, 2015) and Levy (2014b), grant that System 1 attitudes enter into associative, rather than rational, causal transitions with one another, but deny that the states so related are associations rather than more familiar propositional attitudes. Here, however, I'll understand psychological states or attitudes functionally, such that differences of this sort in the very type of causal transitions that attitudes are capable of entering into are sufficient for their being different types of attitude.

[11] People also tend to prefer letters that appear in their names and numbers associated with their birth date compared to other numbers (Kitayama and Karasawa 1997, Nuttin 1985, 1987).

and valence of their implicit attitudes(s). The results of most studies run using the IAT are driven by implicit biases, specifically, rather than other types of implicit attitude.[12]

Some dual-processing findings have been beset by the recent *replication crisis* within social psychology. I'll stay out of the furor, except to note, with Doris (2015: 44–49), that none of this puts the dual-processing perspective generally into doubt. Some sources of evidence may be overturned (many of the studies on semantic priming, in particular, have failed attempts at replication). But at the level of general theoretical framework, the evidence for implicit attitudes is more than overwhelming. Here, I'll simply take the gold standard to be the IAT—probably the best-documented and most well-replicated set of findings (though I'll provide evidence that the attitudes driving these effects are also those responsible for the name-letter and other implicit egotism effects, as well as sympathetic magic).

Few have challenged the IAT results themselves, though some have questioned how large an effect on behavior any individual implicit bias has (e.g., Forscher et al. ms.). Further empirical results will hopefully clarify issues, but conceptually-speaking, this may be the wrong perspective to take in the first place. Implicit associations between concepts form an entire web or network, which might be graphically represented using nodes for concepts and (weights of) edges for (strengths) of association between them (note that these may be positive or negative weightings). We can then understand the pattern of activation and inhibition that any (set of) stimuli lead to in terms of its rippling effects on this entire web. It seems likely that particular sets of associations will form more tightly connected pockets (with higher intra-set weightings) compared to the network as a whole. And we might expect these "pockets" to have more reliable effects on behavior across different types of situation. For example, we might expect BLACK, POOR, UNEMPLOYED, DANGEROUS, and GHETTO to form a particular pocket or "neighborhood" within many American's mental associative networks and for this to have more robust effects on their behavior than any association between BLACK and any other individual concept taken in isolation.

Again, these associative connections between concepts are not reasons-based. The associative transitions that drive the IAT and related dual-processing effects are, instead, essentially governed by the three principles postulated by Hume (1738/1978): resemblance, contiguity, and correlation.[13] We associate likenesses with those people that they're likenesses of because of the likeness (that is, the resemblance between the picture and the person), we associate SALT with PEPPER because they tend to be contiguous or found together in our past experience, and we likely associate many of the concepts in the last paragraph (e.g.,

[12]Implicit biases do not differ from other implicit attitudes in terms of their bare psychological nature (like other implicit attitudes, they are associations between concepts). Rather, these attitudes are implicit *biases* precisely insofar as one wishes to make a further normative judgment about them. This is different than the use of 'bias' in "heuristics and biases," as discussed at the beginning of Ch. 3.

[13]Causation requires the further mysterious "necessary connexion," experimental confirmation of which is of course more tenuous. The connections between implicit associations in this literature and in other strands of the "associative tradition" are complex (cf. Mandelbaum 2015). The open questions include, e.g., whether implicit attitudes in the IAT sense are implemented in connectionist networks (Quek and Ortony 2011 show they can be), and what their relation is to classical conditioning (Olsson et al. 2005 show they interact.)

POOR, UNEMPLOYED, and GHETTO) with BLACK because we've experienced that the properties these concepts track are correlated in our environments.[14]

While implicit attitudes are capable of tracking statistical regularities, they don't conform to—and often conflict with—other normative rules, like those of logic. Perhaps the most striking example comes from studies showing that implicit attitudes are blind to negation. Rozin et al. (1990), for instance, found that participants prefer to drink from a cup labeled "sucrose, table sugar" than from one labeled "not sodium cyanide, not poison," even though they themselves poured normal table sugar into both. Rozin et al. take this to suggest that the causal chain leading to the preference, and hence action, is radically insensitive to even logical negation, one of the simplest rational relations between contents. Participants are averse to the jar in question even though it's labeled "*not* sodium cyanide, *not* poison."[15]

Implicit attitudes' effect on behavior is widespread, including on "microbehaviors."[16] Many of the morally-relevant behaviors influenced by implicit biases are not so "micro," however, and involve serious life decisions like where to move and what career to pursue. This includes implicit racial biases, which are quite widespread (Nosek 2007) and are likely partly responsible for the fact that otherwise identical résumés with "white-sounding" names (like 'Emily' and 'Greg') receive 50% more interviews than those with "black-sounding" names (like 'Lakisha' and 'Jamal') (Bertrand and Mullainathan 2003; see also Dovidio and Gaertner 2000). More unsettling yet, implicit biases raise the likelihood of mistaking tools for guns and shooting those holding them in computer simulations (Payne 2001, Correll et al. 2002). This is supported by convergent work on semantic priming using the *Affect Misattribution Paradigm (AMP)*. For instance, most white Americans rate unknown Chinese language characters more negatively when primed with black faces as compared to white faces (Payne et al. 2005) and mistake pictures of tools for guns more often when primed with black compared to white faces (Payne 2006). AMP scores also correlated with citizens' actual voting behavior in the 2008 US presidential election (Payne et al. 2010).

Implicit biases also influence the severity of criminal sentencing decisions (Blair, Judd, and Chapleau 2004, Eberhart et al. 2006), and likely explain many differences in helping behavior—e.g., the greater likelihood of helping random white compared to black callers in need of a favor (Dovidio and Gaertner 1986).[17] Apart from implicit biases, specifically, Carver et al. (1983) show that participants implicitly primed with "hostility" give (what they

[14] As I argue in Ch. 5, though, these very implicit biases (e.g., between BLACK and DANGEROUS) also *cause* blackness to be correlated with these other properties (e.g., through residential discrimination).

[15] On System 1's characteristic blindness to negation, see also Deutsch et al. (2006), Gawronski and Bodenhausen (2011), Gilbert (1991), and Hasson and Glucksberg (2006).

[16] Dovidio et al. (1997), for instance, find that implicit racial biases predict participants' making less eye contact with and blinking more when interacting with a black interviewer, Wilson et al. (2000) demonstrate that implicit attitudes predict how much participants touch a black confederate's hand, and Bessenoff and Sherman (2000) show that people's implicit attitudes toward the overweight predict how far away they spontaneously sit from overweight individuals. See also Dovidio et al. (2002).

[17] A recent meta-analysis on helping behavior confirms this pattern of results, and shows that it has not subsided over the past 40 years despite decreases in explicit racism (Saucier, Miller, and Doucet 2005). For reviews on implicit biases' effect on morally relevant behavior, see Jost et al. (2009).

believe to be) more intense electrical shocks to another person in a Milgram-style "learning" experiment, and Vohs (2006) shows that participants primed with money (on computer screen savers) behave more selfishly and less helpfully than participants not so primed.

## 2.3 Types of Interaction Between Implicit and Explicit Attitudes

Many psychologists believe that in cases of implicit and explicit conflict or incongruence, it's one's explicit thought that "normally has the last word" (Kahneman 2011: 25). Once you reflect on the bat-and-ball problem, you have no difficulty in overcoming the implicit, System 1 response and giving the correct answer. Many of the studies above, though, show that there are also numerous cases of conflict in which implicit thoughts deliver an output that wins out over or (as I'll put it) "beats" one's explicit, System 2 thoughts in producing behavior. Tamar Gendler (2008a, 2008b), for instance, focuses on implicit attitudes she calls aliefs that are revealed by cases in which they conflict with System 2 responses. Everyone *believes* that the transparent observational platform suspended above the Grand Canyon is perfectly safe, so many people's fear response and considerable trouble walking across it must instead be the product of aliefs. Similarly, one may explicitly believe that a piece of fudge shaped like dog feces is safe and even tasty, so the fact that many people still avoid it (Rozin et al. 1986) is evidence of an alief to the effect that it's disgusting. People are sometimes aware of conflicts between implicit and explicit thoughts and can even feel the tension involved—e.g., in classic cases of weakness of will, often especially when disgust or other gut affective reactions are involved. As Gendler (2014) notes, though, many cases of inter-System conflict revealed by the dual-processing literature may not be phenomenologically apparent.

Disgust, for instance, may operate below the level of awareness—e.g., shaping one's moral judgments (Haidt 2001).[18] Joshua Greene (2001, 2002, 2004) and colleagues have also argued that in response to moral dilemmas, characteristically deontological judgments are produced by implicit affective reactions, whereas characteristically utilitarian moral judgments are made by one's explicit thoughts, a fact that Greene takes to carry weight in the debate between these normative theories. Specifically, it looks as though System 1 is sensitive to whether or not an action uses muscular force: this is why most people have the intuition that it's wrong to push one fat man off a footbridge into the path of a trolley to save five people even though flipping a switch to divert the trolley onto a track with the fat man would not be wrong (Greene et al. 2009, Greene 2013, 2014). Many of these and other dual-processing studies appear to reveal cases in which one's valuings (in Bratman and Doris' sense) are subverted. One's valuings (which haven't been tampered with themselves) are

[18]Wheatley & Haidt (2005) show that participants hypnotized to feel disgust when hearing a particular word make more severe moral judgments about any actions described in a way that contains that word, and Schnall et al. (2008) demonstrate that exposure to artificial "fart spray" and working in a messy room increase the severity of subsequent moral judgments. See also Inbar et al. (2009), who show that disgust sensitivity predicts implicit bias toward homosexuals, as measured by IAT.

brought to bear on the situation and often drive one's explicit judgment, but these responses are overpowered or otherwise "beaten" by one's implicit attitudes in producing behavior. This is one way in which one's implicit and explicit attitudes can interact.

There are other cases in which both Systems come "on-line" but work together: cases in which System 1 gives System 2 some piece of "advice" (e.g., recommends a choice or action) and System 2 checks this answer through its own processing, endorses it, and this then leads to behavior.[19] Again, there are many cases in which System 1 gets things right. But there are other forms of interaction that fall short of outright conflict, as well. Sometimes System 2 is too "lazy" or otherwise overtaxed to check System 1's answer (the "cognitive miser" hypothesis) and other times implicit thoughts mislead one's System 2 processing (e.g., by surreptitiously substituting the answer to some associated question for the one actually asked) (Kahneman & Frederick 2002). For instance, when asked the bat and the ball question, many people may hear "A bat and a ball together cost $1.10. The bat costs $1. [Rather than "The bat costs $1 *more than* the ball.] How much does the ball cost?" In some cases, it may even be normatively better for System 1 to get its way by beating System 2: System 2 is sometimes normatively incorrect.[20] There are, in addition, conflicts that are wholly internal to System 2, as when one has competing explicit desires or even values.

Beating isn't the only type of direct conflict between implicit and explicit attitudes, either. As Doris (2015: 52) notes, implicit attitudes and processes can also *bypass* System 2 entirely. These are cases in which System 1 doesn't need to "win out" over System 2's rejection of its advice because it never alerts System 2 to the fact that there's any decision to be made, such that System 2 is never brought "on-line" in the first place.[21]

Bypassing can sometimes lead to *post-hoc rationalization*. This occurs when choices or behavior are made on the basis of some System 1 response (e.g., a disgust reaction) that one is unaware of, and which one explicitly tries to come up with a justification for after the fact (often because prompted to do so). In the classic experiment, participants asked to choose the highest quality pantyhose from four identical pairs are much more likely to choose the pair on the right, and offer various explanations of their behavior—none of which, however, include its being on the right. Indeed, participants explicitly deny that they take spatial position to provide any reason for their choice whatsoever (Nisbett and Wilson 1977). This seems to be a case of System 2 having to come up with a justification after the fact precisely because it was unaware of the basis on which System 1 actually made the choice.[22]

Failure of rationalization can still lead to moral "dumbfounding," when people persist in a judgment or choice even after they can come up with no reason for it (see, e.g., Haidt 2001 and Cushman et al. 2006). Dumbfounding may bleed into a type of beating. People insist

[19]Railton (2009, 2011, 2014) gives these cases considerable attention.

[20]Gerd Gigerenzer (1999, 2007) is perhaps the foremost champion of System 1 in this sense.

[21]Of course, bypassing still "interferes" with one's dispositional, standing System 2 values; it's just that these values aren't occurrently brought to bear on the situation at hand, as they are in cases of beating.

[22]If System 2 had come on-line before the choice and assessed the situation or problem itself, one presumably wouldn't try to rationalize the choice, but would instead recognize the discrepancy as a case where one had "given in" to temptation, disgust, or the like. This would then be a case of beating.

that sibling incest and masturbating with a dead chicken, e.g., are morally wrong, even after the examples at hand have been stipulated to involve no (possible) harm, such that the risk of harm can't be used as a justification. The disgust reaction to these actions nonetheless continues to win out in producing one's judgment—one doesn't revise it even after System 2 fails to come up with any justification for it. It might be objected that participants in these studies are taking their disgust to provide a reason for moral condemnation (Jacobson 2012), but this seems unlikely. Especially given participants' repeated attempts to provide what they'd take to be a justification for their dumbfounded judgments, it seems that*from their own point of view*, disgust provides no reason for moral condemnation.

In sum, not all cases of inter-System conflict in which System 1 wins out threaten autonomy (System 1 often gets things right). But many do. Still, the types of conflict discussed so far don't fully account for the eeriness of many dual-processing results. Beating and bypassing interfere with autonomy, but it's not clear that they do so in any terribly unique way—anymore, say, than strong temptations might overpower one's better judgment or the way a brain tumor might bypass System 2 processing entirely. Neither of these types of inter-System conflict do justice to the sense in which System 1 can seem *manipulative*—cases like the Hodson et al. (2002) findings where System 1 infects the very System 2 capacities that many think are central to autonomy. In the next section, I'll fit such cases of "hijacking" into the taxonomy of inter-System interactions begun in this section.

## 2.4 Hijacking Reason

*Hijacking*, as I'm calling it, completes the framework of ways that System 1 can affect behavior, based on its interaction with System 2. A more complete framework might include ways that System 2 can affect behavior based on its interaction with System 1, but the former is enough to occupy us here. The possibility of other variations already noted, then, we can roughly divide the ways that System 1 can affect behavior into four broad categories:[23]

- *Advising*: System 1 suggests a behavior or other output to System 2, which then reviews and accepts it—e.g., deciding that because one just "feels" more like moving to Georgia (than, e.g., Alaska), one will move to Georgia (presumably, because everything else is equal and one is otherwise indifferent between the two courses of action).[24]

- *Beating*: System 1 produces an output that competes with System 2's output for control over behavior, and System 1 wins out (perhaps after unsuccessfully attempting to advise System 2 to take this course of action).[25]

[23]Doris (2015: 53–61) and Smith (2015) produce similar classificatory schemes, but crucially leave out hijacking. I've already noted that Doris often comes close, though. Smith (2015: 191–3) also explicitly compares the dual-processing results to intentional hypnosis by another agent, but neither author draws the distinction between hijacking and bypassing central to my view here. For more on both positions, see Ch. 4.

[24]Doris (2015: 55) notes that associative "advice" is most likely to be accepted for "tie-breaking."

[25]Of course, there are cases where System 2 wins, as well—successful executions of willpower.

- *Bypassing*: System 1 produces an output that would conflict with System 2, were the latter system operative (with respect to a particular practical question or decision), but in fact System 1 prevents System 2 from ever coming on-line in the first place.[26]

- *Hijacking*: System 1 influences behavior *by* influencing System 2's own processing, which is the proximate cause of behavior—i.e., System 1's influence on behavior is mediated by first producing a System 2 response that conforms to it, and so ultimately leads to System 1-concordant overt behavior. For example, one might come up with various rationalizations for moving to Georgia—e.g., that the job there really is superior, or that warm weather is more important than adventure or proximity to family—that occur before one makes a decision, and determine how one makes it.

Two clarifications about the notion of hijacking. First, it still involves a type of bypassing. Crucially, though, this isn't the sort of bypassing of System 2 in its entirety that others have had in mind. Doris (2015: 72–3) writes, for instance, that it's "[b]est to distinguish cases where happenstance engages rational capacities from cases where happenstance bypasses rational capacities. I've been worried about cases of bypassing: influences that are not vetted by rational capacities. The mediating saliences we're now considering may be cast as cases where rational capacities are engaged... I've no stake in denying the existence of such cases, even many such cases. But this does not rule out the existence of bypassing, where reason does not get engaged."[27] In cases of hijacking, however, one's rational capacities precisely *are* engaged in vetting one's response. This is why they seem to present such an acute threat to autonomy: they take some of the very capacities that many thought were central to autonomy away from our selves. That said, hijacking does seem to threaten autonomy because it's a type of "bypassing"—not of System 2 entirely, but of one's *valuings*. In cases of hijacking, the System 2 capacities of practical deliberation are wrested out of their "normal" control by one's valuings and driven by System 1 instead. Thus, while important parts of System 2 precisely are not bypassed, one's valuings may be (at least to some extent).[28]

[26]Many authors suggest that it's a System 1 process that brings System 2 online to address any particular question (or not, in cases of bypassing) (Kahneman and Frederick 2002). In addition, *System 2* sometimes "bypasses" *System 1*—e.g., through the types of strategy employed to avoid tempting situations in the first place. Again, a more complete taxonomy might include these types of inter-System interaction.

[27]On the same page, though, Doris (2015: 72) does note something akin to hijacking: "the cognitions that pattern motivation may be no less subject to capricious variation than the motivations themselves... now, the problem concerns rationally arbitrary influences on the saliences that help structure preference."

[28]Washington and Kelly (forthcoming) witness the need for this distinction: they note that it's implausible that people's deliberative capacities and conscious reflection are "completely bypassed" in many of the relevant findings. This leaves open that System 2 might be bypassed but not "completely" (rather, only certain parts of it)—even if one's practical reasoning and judgments are in use (being hijacked), one's valuings, specifically, might still be bypassed. When Nahmias and Murray (2010) and Murray and Nahmias (2014) introduce the notion of "bypassing" adopted by Doris (in a somewhat different domain), they simply don't have these distinctions in mind. But because "bypassing" seems to have come to mean "complete bypassing" in the implicit attitudes literature, there's need to distinguish this from what I here call "hijacking."

Second, I've suggested that when hijacking occurs, one's practical reasoning and deliberation is rendered (at least in relevant parts) into mere rationalization. This is different, though, than the type of *post-hoc* rationalization or confabulation discussed at the end of Section 3. In cases of post-hoc rationalization, one engages practical reasoning to come up with some justification after the fact, even though one's action had already been determined by other (System 1) sources. But many of the dual-processing results seem to unearth a similar type of rationalization that instead occurs *before* one's decision—reasoning that's tailor-made before the decision, and which determines it, precisely to System 1's specifications (assuming, of course, that the hijacking of one's practical reasoning and deliberation is successful). The remainder of this section reviews the evidence for hijacking in this sense.

Dovidio and Gaertner (2000) asked participants to provide recommendations for a peer counseling program on the basis of interview excerpts with candidates. These excerpts varied the candidate's race (black vs. white) and strength of qualifications (clearly strong, ambiguous, or clearly weak). Whether the candidate was black or white didn't have any effect on the strength of participants' recommendations of clearly strong or clearly weak candidates. However, among candidates with ambiguous credentials, white candidates were recommended over 20% more than black candidates. Whites' ambiguous qualifications are treated as strong, whereas when blacks have these same qualifications, they're treated as weak. This appears to be a case where one's implicit biases are clearly impairing, rather than promoting, one's reasons-responsiveness (specifically, one's assessment of or sensitivity to the evidence)—cases of rationalizing one's judgments (about the qualifications of the candidates). Tellingly, System 1 plays its hand only when it won't be noticed—not, that is, when the candidates' qualifications are clear enough that any rationalization might be detected by System 2 (in the clearly strong and clearly weak qualification conditions).

The findings discussed in Section 1—Hodson, Dovidio, and Gaertner (2002)—support this interpretation. Participants were asked to assess prospective students for college admission who either had strong high school grades and SAT scores, only strong grades, or only high SAT scores, and who were either white or black. For lower-prejudice participants, black applicants were recommended more strongly across conditions. For higher-prejudice participants, there was no significant effect on admissions recommendations for the strongest- and weakest-qualified applicants, but for those with mixed qualifications, white applicants were recommended more strongly than black applicants. Moreover, higher-prejudice subjects claimed to place more importance on whichever of the credentials the black candidate *did not have* for college admission generally. That is, when a black candidate had only high SAT scores, these participants took strong grades to be more important for admission, but when the black candidate had strong grades, they instead took high SAT scores to be more important for admission.[29] This is an even clearer instance of *pre-hoc* rationalization—one's very practical deliberation about the general balance of reasons is hijacked by one's implicit

[29]The measure of prejudice used in these studies was Brighams (1993) Attitudes Toward Blacks Scale, technically a measure of explicit racism but which Hodson, Dovidio, and Gaertner (2002: 462) explain functions in a way that likely tracks implicit racial biases in this context.

racial biases. Note also that it's not clear what *other* reasons these participants might even try to adduce for their overt behavior—their admissions recommendations—besides their assessment of the candidate's qualifications or credentials and their relative importance.

Son Hing et al. (2008) show that participants who display IAT biases against Asians are more likely not to recommend Asians for jobs when there's a non-race-related excuse not to hire them (ambiguity of how well the applicant's qualifications fit the job), compared to when there is no such excuse. This effect isn't found for those with low racial bias on the IAT nor, more importantly, for white job applicants. Again, this strongly suggests that participants are making the choices they do *because* of their implicit racial biases, and that they then rationalize these choices with their assessment of the candidate's credentials, qualifications, or "fit" with the job. These rationalizations, in turn, determine one's choice and behavior.

Another possibility is that one's implicit attitudes are merely a common cause of both overt behavior (e.g., admissions or hiring decisions) and the wayward System 2 reasoning (e.g., assessments of merit), but that implicit attitudes do not affect one's behavior *through* their effect on one's reasoning. (Or, in some cases of post-hoc rationalization, it may even be that System 2 reasoning is caused by one's behavior, which in turn is directly caused by System 1 attitudes.) Some of the studies discussed here don't rule out this possibility, but others do. Chaxel (2015), for instance, shows that the effect of IAT scores on recommendations about which professor to give a performance bonus to are fully mediated by distortions in the relative importance that participants assign to different attributes (involving service, research, and teaching credentials). In other words, after taking into account the effect that implicit biases have on decisions and behavior *through* their effect on System 2 reasoning about the relative importance of the values involved (how to weight the different credentials against each other), implicit biases have no significant additional (direct, unmediated) effect on decision and behavior. This is strong evidence of hijacking.

Uhlmann and Cohen (2005) asked participants to assess candidates for a police chief or women's studies professor job opening. The candidates were described as being either male or female and, in the police chief study, being either "streetwise" or "formally educated"; in the professor study as being either "purely academic" or "activist." Uhlmann and Cohen found that female participants took being an activist to be a more important qualification for being a women's study professor *when* female candidates possessed this attribute, but not when males did. Similarly, male participants took being formally educated (and being family oriented) to be more important qualifications for the job of police chief when male candidates possessed them, but if anything, rated these same qualities as less important when they were possessed by women. In these sorts of case, the only reasons one has to base any evaluation or hiring recommendation on are presumably one's (comparative) assessments of the candidates' merit. But it seems that these assessments—what one takes to constitute merit, and so reasons to hire candidates—are mere (pre-hoc) rationalizations.[30]

[30] As Ullman and Cohen (2005: 479) put it, participants in their experiment "tailored their criteria to favor whatever qualities the individual applicant of the desired gender happened to have," which these authors take to constitute "a novel and pernicious source of discrimination: definitions of merit designed to fit the

It might be argued that some of the results above are driven by motivated reasoning of the already-familiar variety—tailoring one's descriptive beliefs or theoretical reasoning to a desired conclusion (perhaps to conform to the descriptive beliefs one already holds)—rather than any more direct interference with *practical* reasoning specifically. And a number of studies do demonstrate that implicit attitudes can produce this type of "epistemic hijacking" or "hijacking of theoretical reasoning."[31] Other findings can't be chalked up to motivated reasoning of this traditional epistemic or theoretical variety, however.

Uhlmann, Pizarro, and Ditto (2009) gave participants a version of the trolley task, but with a twist. Half of the participants were asked if they would push 'Tyrone Payton' off a footbridge into an oncoming trolley in order to save 100 members of the New York Philharmonic (recall Bertrand and Mullainathan 2003); the other half were asked whether they would push 'Chip Ellsworth III' off the footbridge to save 100 members of the Harlem Jazz Orchestra. Interestingly, liberals but not conservatives were more likely to offer consequentialist justifications when this would involve sacrificing Chip compared to Tyrone (a result replicated in another experiment using a "life boat"-type scenario). In contrast, conservatives but not liberals were more likely to endorse consequentialist justifications of "collateral damage" for a military strike when the innocent civilians who'd be killed were Iraqi rather than American. Thus, even moral reasoning—quintessential practical reasoning—can be influenced in quasi-goal-directed ways by implicit biases. These even affect when one thinks consequentialist (rather than deontological) reasoning is and is not appropriate.

As Uhlmann, Pizarro, and Ditto (2009) note, their findings thus "provide evidence that motivation can influence not only our descriptive beliefs about how the world is (Dunning & Cohen 1992, Kunda 1987, Norton et al. 2004, Simon 2004, Simon et al. 2004, Uhlmann & Cohen 2005) but also prescriptive beliefs about how the world *ought* to be."[32] Uhlmann et al. also explicitly contrast this type of rationalization with the *post-hoc* variety, writing that they "observed how selective endorsement of moral principles emerged to support those conclusions" drawn by one's implicit attitudes. In other words, these results demonstrate specifically *practical* hijacking, in which one's motivations and normative judgments—not just descriptive beliefs or theoretical reasoning—are wrested from the control of one's values.

The results of the studies above and a host of others establish that cases of hijacking do exist: cases in which System 2 practical reasoning is not overpowered or left out of the loop leading to choice and action, but is instead *used* as a "means" to the "ends" of one's implicit attitudes. In these cases, System 1 co-opts System 2 practical reasoning for its own purposes, turning its deliverances into mere rationalizations—rationalizations which then cause one's decisions and behavior rather than tagging along after it. In the limiting cases, the *only*

idiosyncratic qualifications of applicants who belong to favored groups."

[31]On motivated reasoning, see Kunda (1990) and Brownstein (2003). Ditto & Liu (2011) and Liu & Ditto (2013) find that implicit moral attitudes can hijack descriptive beliefs. Johnson et al. (1995) show that introducing evidence deemed inadmissible in court is more likely to lead to the conviction of black compared to white defendants (cf. Faranda and Gaertner 1979, Hodson et al. 2005). Compare this inability to ignore undermined evidence with the blindness to negation in Rozin et al.'s (1990) "*not* poison" study.

[32]See also Ditto, Pizarro, and Tannenbaum (2009).

reasons potentially available to explain one's action may be pre-hoc rationalizations.

Hijacking is thus where the "manipulativeness" of implicit attitudes truly comes to the fore. Beating and bypassing may threaten autonomy, but as noted above, it's not clear that the ways they do so are that theoretically novel. The effects of beating and bypassing are seemingly comparable to those of, say, a(n unintelligent) brain tumor. A brain tumor (or irresistible, compulsive urge or desire) might beat or bypass System 2, as well, and that might well threaten autonomy. But these threats are not *sinister* in the same way that, e.g., the Hodson, Dovidio, and Gaertner (2002) findings are. A brain tumor might affect your admissions recommendations; it might even make you favor a particular race or qualification more. But it's hard to imagine a brain tumor being "intelligent" enough to make you reason that SAT scores are more important than grades *when* the black candidate has better grades, but also to make you think the reverse—that grades are more important than SAT scores—*when* the black candidate has better SAT scores. *That* is downright diabolical.[33]

By analogy: To stop, sabotage, or otherwise get around or ground an airplane is one thing, but to *use* the plane for your own purposes, you actually have to *know how to fly it*. For the former, any well-placed wrench, explosive, or gremlin will do. In stark contrast, *hijacking* the plane requires a further sophistication and appreciation for how it works *qua* airplane. So, I say, for the ways that System 1 can affect behavior, based on its interactions with System 2. To beat or bypass System 2, it needs certain characteristics, but these needn't be terribly fancy. But *hijacking* requires precisely the sort of sophisticated "intelligence" of how practical reasoning works that seems to license talk of implicit biases' "goals," "intentionality," or even "manipulativeness." We've now seen how hijacking is the most important type of inter-System interaction for bringing out this intuition, but we've yet to clarify the sense in which it really counts as (quasi)-intentional, a task to which we turn in the final section.

## 2.5 Hijacking as Decrease in Counterfactual Dependence

As mentioned in Chapter 1, a number of experimental philosophers have begun to investigate why manipulation by another individual agent intuitively undermines autonomy, which I discuss below. Several authors have suggested that their results are best explained by interventionist theories of causation (Murray & Lombrozo 2017 and Deery & Nahmias 2017), which have recently become popular in philosophical accounts of mental causation (e.g., Campbell 2010, Ismael 2013, and Roskies 2012) and in psychological accounts of people's intuitive judgments about it (e.g., Gerstenberg & Lagnado forthcoming, Lagnado & Channon 2008, Lagnado, Gerstenberg, and Zultan 2013, Lombrozo 2010, and Sloman 2005).

[33]Of course, if a brain tumor did have the same sort of fine-grained, quasi-intentional control over your cognitive economy, I think it'd be just as spooky as hijacking. I don't mean to rule out this possibility, only to contrast the dual-processing results with what seems far more probable for your typical brain tumor.

Interventionist theories of causation (e.g., Spirtes, Glymour, and Scheines 1993, Pearl 2000, Hitchcock 2001, Woodward 2003, Woodward & Hitchcock 2003, and Hitchcock & Woodward 2003) use causal models to represent the relationships of counterfactual dependence among different variables. The values of these variables in turn represent events—e.g., $X$=1 if the gate is open; $X$=0 if the gate is closed. A causal model consists of a set of structural equations specifying the counterfactual dependence relations between the variables in the model, represented by directed arrows. These equations represent counterfactuals of the form: "If it were the case that the value of $X$ was $x$, the value of $Y$ was $y\ldots$ and the value of $Z_n$=$z_n$, then the value of $Y$=$\int(X,\ Y,\ \ldots Z_n)$, in background conditions $C$" (where $X$, $Y$, and $Z$ represent the variables in the model; $x$, $y$, and $z$ represent the respective values of these variables; and $C$ represents any direct causes of $Y$ not explicitly represented in the model). A causal statement is *invariant* if it describes how $Y$'s value would change under a wide range of interventions on the value of $X$—i.e., if it is *counterfactually robust*.

It's appropriate to count $X$ as a cause of $Y$ if $X$ and $Y$ are correlated under (hypothetical) interventions on $X$, where $I$ is an *intervention* (variable) for $X$ with respect to $Y$ if it meets the following conditions (Woodward and Hitchcock 2003: 12–13):

1. $I$ is causally relevant to $X$.
2. $I$ is not causally relevant to $Y$ otherwise than through $X$.
3. $I$ is not correlated with any $Z$ causally relevant to $Y$ otherwise than through $X$.
4. $I$, when it takes certain values, screens off any influence on $X$ by other (usual) causally relevant variables.

In other words, we perform an intervention on $X$—altering its value while holding fixed those of independent variables, and in a way that "surgically" screens off $X$ from the influence of any other variables—and see if this leads to a systematic change in the value of $Y$. If so, $X$ is a cause of $Y$. This is the sort of test that constitutes the experimental gold standard in the sciences—the sort of test one might perform, for instance, in a clinical drug trial.

Using such models, we can also account for more specific differences in *how* causally relevant certain variables are to other variables—that is, we can compare the relative counterfactual dependence of an outcome on its different causes. Simplifying (see Woodward 2003: Ch. 6), the dependence relation, $R_1$, between $X$ and $Y$ is *stronger* than the dependence relation, $R_2$, between $Y$ and any of its other causes, $Z$, just in case:

1. Holding fixed background conditions $C$, $R_1$ predicts the value of $Y$ under a wider range of interventions than $R_2$.
2. $R_1$ predicts the value of $Y$ under a wider range of changes to background conditions $C$ than $R_2$.

Intuitively, when conditions 1 and 2 hold, the counterfactuals encoded in $R_1$ are more robust—holding in a wider range of counterfactual situations—than those encoded by $R_2$. Thus, the variations in the value of $Y$ are more sensitive to the value of $X$ than the value of any other variables in the widest range of relevant background conditions.

Deery & Nahmias (2017) use this framework to further define a notion of causal sourcehood: $X$ is the *causal source* of $Y$ iff $X$ bears the *strongest* invariance relation to $Y$ (among $Y$'s causes). They argue that in cases of manipulation, the manipulator's mental states (e.g., her desires or intentions), rather than those of the manipulee, come to bear the strongest invariance relation to the outcome the manipulee is manipulated to perform.

Typically, teleological or goal-directed events bear stronger invariance relations with their outcomes than non-intentional (and unintentional) events, since plans tend to be achieved despite variations in the context and means needed to bring them about. Agents with goals "find a way" in the face of obstacles. As James (1890: 20) puts it:

> Romeo wants Juliet as the filings want the magnet; and if no obstacles intervene he moves towards her by as straight a line as they. But Romeo and Juliet, if a wall be built between them, do not remain idiotically pressing their faces against its opposite sides like the magnet and the filings [when a card or other barrier is placed between them]. Romeo soon finds a circuitous way, by scaling the wall or otherwise, of touching Juliet's lips directly. With the filings the path is fixed; whether it reaches the end depends on accidents. With the lover it is the end which is fixed, the path may be modified indefinitely.[34]

As with iron filings, so with Paris if he somehow "pursues" Juliet unintentionally, and with our friend the brain tumor. The dependence relation between the outcome and Romeo's mental states is more robust, and less sensitive to variations in background conditions and contingent obstacles in the causal chain between them, compared to these variables.

Drawing on Lombrozo's (2010) *exportable dependence theory*—an interventionist theory of causal attribution—Murray & Lombrozo (2017) suggest that in general, people believe that the stronger the invariance relation between one factor and an outcome, the weaker the invariance relation is between other factors in the same causal chain and that outcome. (Intuitively, the more variance in an outcome's occurrence one factor accounts for, the less variance any other factor can account for.) Thus, in cases where another factor in the causal chain is not only a human being with the goal of bringing the outcome about, but also of doing so *through* the manipulee's own actions (and mental states), the invariance relation between the manipulee's (non-manipulated) mental states and the outcome will be especially weak (much weaker, at least, than when the prior causal factor is not the intentional action of a manipulator but the behavior of magnetized iron filings or a brain tumor).

[34]Cf. Heider (1958) and Lombrozo (2010). There are of course exceptions. Agents acting intentionally might forget the means of bringing their goals about or have them otherwise interfered with, and some non-intentional factors (like paralysis) can in some cases bear very strong invariance relations with outcomes (in this case, especially omissions). Thanks to R. Jay Wallace for bringing cases of this kind to my attention.

Murray & Lombrozo (2017) ask participants about causal chains comprised of a first variable (the "manipulator") influencing a second variable (the "manipulee"). Six experiments independently vary the effect of the first actor's (i) simply being an agent, (ii) acting intentionally, (iii) foreseeing the effects of her actions, (iv) intentionally causing the manipulee's action (but not the outcome it leads to), (v) intending the outcome as well, and (vi) intentionally producing the outcome by causing the manipulee's action by bypassing her mental states (e.g., by using hypnosis or mind-controlling drugs). What they find is a steady decrease in attributions of autonomy to the manipulee through conditions (iv)-(vi), with bypassing presenting the largest threat. These results appear to confirm that manipulation is intuitively threatening to autonomy because it involves a decrease in the counterfactual dependence of the outcome on the manipulee's mental states (as I'll go on to argue, her values, specifically). Bypassing presents the limiting case, in which some of the manipulee's mental states are cut out of the causal chain leading to her action entirely. But the results also suggest that this is just one extreme of a more general effect: in conditions (iv)-(vi), as the manipulator progressively intends more events in the causal chain leading up to the outcome (increasing its counterfactual dependence on her intention), and the outcome's counterfactual dependence on the manipulee's own mental states thereby decreases, the manipulee's perceived autonomy over bringing about the outcome progressively decreases, as well.

In sum, the threat to autonomy presented by the causal influence of other agents seems to be a graded effect, corresponding to the strength of the invariance relations involved. Intentional influence is especially threatening because it involves especially strong counterfactual dependence (and inversely proportional weakening of the relation between the manipulee's mental states and the outcome). These are quantitative differences in degree (as Murray & Lombrozo 2017 show), and in this sense, the threat that manipulation presents to autonomy is just one extreme along a graded spectrum of threats: those of counterfactual dependence. As Deery & Nahmias (2017) point out, though, manipulation often also involves a qualitative shift in the *relative* degree of counterfactual dependence. At the point at which the manipulator's mental states (e.g., her values) come to bear a *stronger* invariance relation with the action and outcome than the manipulee's, the causal source of the outcome switches, and this may be a particularly important milestone for questions of autonomy.

## 2.6 Conclusion

Return now to hijacking. Implicit associations may not bear invariance relations with their outcomes (the effects they have on one's actions or reasoning) as strong as the invariance relations of the explicit System 2 intentions of another agent, but in cases of hijacking (and some cases of bypassing), they do bear stronger invariance relations to those outcomes than completely inanimate factors typically would. If Juliet is behind a wall, Romeo will climb; if she's through a tunnel, he'll crawl. The high prejudice participants' implicit biases in the Hodson, Dovidio, and Gaertner (2002) study display eerily similar behavior: if placing more weight on high SAT scores would bar the black candidate from admission, then these partici-

pants weight high SAT scores more heavily; if placing more weight on good grades would bar the black candidate from admission, they claim grades are more important. All unwittingly. Just as Romeo takes whatever means will realize his goal of kissing Juliet, one's implicit associations produce whatever weighting of criteria will lead to denying admittance to black candidates.[35] It's just that implicit associations shift these weightings quasi-intentionally or as a quasi-means to realizing the "goal" of discriminating against black people.

Interventionist theories of causation allow us to cash out the metaphors: "Quasi-goal-directed," "quasi-intentional," and their cognates can all be understood simply in terms of the strength of invariance relations between causal variables. Being intentionally manipulated by another agent presents a unique threat to autonomy because of its uniquely strong counterfactual dependence or invariance relation. Again, hijacking exhibits a similar threat because it involves System 1 implicit associations preferentially bringing about outcomes through counterfactual relations of intermediate strength—considerably stronger than those of inanimate causal transitions, like those involving iron filings and brain tumors, but not as strong as those exhibited by literal manipulators. This, I suggest, explains the intuitively "manipulative" sense evinced by many implicit bias and other dual-processing results, allowing us to understand the force of the analogy without recourse to treating System 1 like any kind of homunculus with its own literal goals, intentions, or other teleological states.[36]

Further results support the claim that implicit associations can enter into invariance relations of intermediately high degree. Glaser & Kihlstrom (2006) suggest that System 1 is able to unconsciously monitor its own operations for potential departures from processing "goals" and correct for deviations.[37] And hijacking suggests that System 1 is also able to monitor System 2 processing and correct for potential departures from its (System 1's) "goals." Indeed, six studies by Marien et al. (2012) demonstrate as much. They find that subliminal priming of implicit System 1 goals impairs performance on tasks that support unrelated goals—specifically with respect to *executive functioning* (in this case, the inhibition of prepotent responses and textual error detection). Executive functioning involves regulatory processes that hold goal-relevant information in working memory (like the means needed to bring the goal about), inhibit competing goals and other potentially interfering processes, and monitor goal pursuit, processing feedback in order to adjust ongoing behavior. Executive functions are taken to have limited capacity (Kahneman 1973, Navon 1984, Pashler 1998). Thus, Marien et al.'s results suggest that implicit System 1 "goals" draw on the same executive functioning capacities as explicit System 2 goals (such that priming the former takes up some shared limited capacity and impairs performance on the latter). As Marien et al. (2012: 399) themselves put it, this demonstrates "that executive functions are

[35] "Ends" encoded by implicit associations may be less determinate than those of explicit intentions.

[36] It's become ubiquitous in the social psychology literature to speak of implicit "goals" and "intentions." I think we should be wary of this language, especially because it's not clear that associations *could* count as literal intentions, since they have associative, rather than propositional, structure. On my view, what's central is (degree of) counterfactual dependence, not whether the states count as intentions or goals.

[37] The primary evidence comes from work on reverse priming (e.g., Glaser & Banaji 1999, Glaser 2003).

being "hijacked" upon the activation of unconscious goals."[38]

These results support a quantitative version of the analogy to manipulation: System 1 is "smart" enough to "know" how to engage many of the same processes as System 2 attitudes—e.g., to engage the monitoring and regulatory processes required to pursue the "means" to an "end," and how to fit means to ends. Moreover, System 1 attitudes know how to hijack System 2's own executive functions, wresting them away from the control of System 2 values. Again, it's in this sense that hijacking seems creepier than bypassing and beating, and establishes something about System 1 that these phenomena do not. Hijacking doesn't work around or overpower but instead operates the very machinery of System 2 for its own "purposes." You wouldn't expect this sort of behavior—"knowing" which attributes to weight more heavily or engaging executive functions—out of a typical brain tumor. Thus, hijacking displays that System 1 attitudes are smarter than we'd otherwise have evidence for—that they know how to "fly" System 2. More precisely, hijacking displays that System 1 attitudes can stand in stronger invariance relations with their outcomes than brute, inanimate causes.

Many results also support a qualitative analogy to manipulation: they demonstrate a change in causal sourcehood in Deery and Nahmias' (2017) sense. Much of the work on differential weighing of criteria, like the Hodson, Dovidio, and Gaertner (2002) results, seems to demonstrate cases where one's action bears a *stronger* invariance relation with one's implicit associations *than* one's explicit System 2 valuings.[39] Galdi, Arcuri, and Gawronski (2008) investigated this type of possibility in the field. They asked people from Vincenza, Italy whether a U.S. military base there should be enlarged or not (a controversial, and genuine possibility in Vincenza during the data collection in 2007). At $t_1$, one week before the decision, participants were given the *choice* about whether they were in favor, undecided, or against the enlargement; a 10 question survey on their value-relevant *evaluative/normative beliefs* about the "environmental, political, economic, and social consequences of enlargement"; and an IAT measuring their *implicit attitudes*, asking them to categorize pictures of the military base with positive and negative words. All measures were then completed again at $t_2$, one week later. For participants who were initially undecided, implicit associations at $t_1$ significantly predicted evaluative/normative beliefs at $t_2$, an effect the authors liken to the type of hijacking stressed in this chapter. In addition, and more intriguingly, implicit associations at $t_1$ were *better* predictors of people's *choices* about the military base at $t_2$ than their evaluative/normative beliefs at $t_1$ (prior IAT scores had a significant effect on choices; prior evaluative/normative beliefs did not). In other words, these results provide a direct real-world demonstration of implicit associations bearing a stronger invariance relation with eventual deliberation and choice than one's explicit evaluative/normative beliefs.[40]

[38]Further evidence that implicit attitudes can occupy executive functioning is discussed in Dijksterhuis & Aarts (2010), Hassin et al. (2009), and Suhler & Churchland (2009). To my knowledge, the only other work using the term "hijacking" in the context of implicit attitudes or dual-processing relates to the emotions, in particular emotions' role in addiction (see Goleman 1995, Bechara 2005, and Levy 2013).

[39]One recent meta-analysis reveals that several types of discriminatory behavior depend more on System 1 implicit biases than System 2 explicit attitudes (Greenwald et al. 2009), though see Oswald et al. (2013).

[40]Galdi, Arcuri, and Gawronski (2008: 1101) write that "these results suggest the possibility that fu-

In hijacking, System 2 practical reasoning and reflective deliberation are themselves wrested away from the control of (made less counterfactually dependent on) one's self: in particular, they're made less dependent on what one explicitly cares about, values, and identifies with. This is a more insidious threat than beating and bypassing. When the latter occur, one typically notices that one's action or decision is not what one explicitly deliberated and judged it should be. In contrast, when one's practical reasoning and deliberation is engaged but manipulated, one can have every sense that one *is* in control. We've now seen ample evidence that hijacking occurs in real-time. But we have not seen the most impressive evidence. In the next chapter, we cross campus into the labs of behavioral economics, findings from which situate implicit biases within the larger stable of System 1 attitudes and also show just how deep and long-lasting the threat to autonomy that hijacking presents is.

ture decisions of undecided individuals can be predicted by measuring their current automatic associations [Arcuri et al., 2008]. Specifically, the available results [Gawronski, Geschke, and Banse, 2003; Hugenberg & Bodenhausen, 2003] suggest that automatic associations could distort the processing of new information (e.g., by means of selective processing or biased interpretation), such that future decisions that are based on such distorted information will be in line with previously existing automatic associations... Hence, an individual may develop a conscious preference for candidate A over candidate B over the course of *deliberating* about the two options, which is rooted in the biasing influence of automatic associations on the processing of new information. From this perspective, the ultimate decision of an undecided individual may be determined, in a more or less probable sense, long before this individual consciously endorses a preference for one candidate over the other." Note also that among participants who were already decided at $t_1$, explicit evaluative/normative beliefs at $t_1$ predicted not only their choice at $t_2$ but also implicit associations at $t_2$.

# Chapter 3

# Framing THE GOOD

## 3.1 Introduction

This chapter extends the argument of Chapter 2 with a different set of dual-processing findings—those from behavioral economics. Chapter 2 argued that results from social psychology call our ordinary conception of ourselves as autonomous agents into question in numerous situations. In some cases, our implicit attitudes hijack our very processes of practical reasoning and deliberation, which many think are central to autonomy. In this chapter, I use the behavioral economic findings to show just how deep and persistent a threat hijacking presents. In some cases, one's preferences and very conception of THE GOOD can be hijacked. Implicit attitudes can even interfere with what one thinks one values.

Some results from social psychology already attest to this fact. Gardner et al. (1999), for instance, asked participants to read a story and circle all the pronouns contained therein, and to then complete a "values questionnaire." In one condition, the story included first-personal singular pronouns (e.g., 'I', 'mine') and in another condition plural pronouns (e.g., 'we', 'ours'), such that the story participants read was about a trip that one either took alone or as part of a group. Those who read the story with plural pronouns were more likely to say that "collectivist" properties (e.g., belongingness, friendship, family safety) were "guiding principles in their lives," whereas those who read the story with singular pronouns were more likely to self-attribute "individualist" principles (e.g., freedom, independence, and choosing one's own goals). This *pronoun effect* has also been shown to affect moral behavior—e.g., the likelihood of offering a bribe.[1] Plausibly, people are being led astray about what their own values are in these cases—what they think they value isn't actually what they value (the pronouns haven't literally changed the "guiding principles in their lives"), and in a way that affects behavior. The behavioral economics literature corroborates this interpretation.

The usual gloss on the behavioral economic findings of the past several decades takes them to show that agents often violate the axioms of rational choice theory and decision

[1]For a recent meta-analysis of priming individualist and collectivist attitudes, see Oyserman and Lee (2008). On the pronoun effect, see also Brewer and Gardner (1996: 89-90) and Gardner et al. (2002).

theory (expected utility theory)—the axioms these theories rely on to explain behavior. Because conforming to these axioms is taken to be (partly) constitutive of what preferences *are*, the findings naturally suggest a picture in which one's preferences are bypassed: in which something else is working around and driving action instead of them.

As before, I argue that while preferences can be bypassed, the more sinister threat to autonomy is presented by cases in which one's preferences are hijacked by contextual "framing" effects, much as I argued explicit evaluative/normative belief and reasoning can be hijacked in Chapter 2. Indeed, some results show that even whether one sees something as good or bad—one's very conception(s) of THE GOOD and THE BAD—can be influenced by factors other than one's valuings, factors that one wouldn't take to provide any reason or justification whatsoever. According to growing consensus within behavioral economics, it's not so much that one's preferences—which represent one's conception of THE GOOD—are bypassed, but that one's preferences are initially "constructed" in a way that's highly sensitive to factors that one does not oneself value. Preferences are lasting, stable dispositions, and they affect what other, future preferences one can rationally form. This adds diachronic bite to the threat introduced in Chapter 2: implicit attitudes can have deep and lasting effects on explicit System 2 processing of all sorts, even on what one (thinks one) values.[2]

## 3.2 Dual-Processing: Behavioral Economics

Like the work in social psychology discussed in the last chapter, behavioral economics grew out of the groundbreaking work of Tversky and Kahneman, but since then these two bodies of research have proceeded somewhat in parallel. Before discussing the results specific to behavioral economics, a quick canvassing of the classic heuristics (or System 1 "shortcuts") that Tversky and Kahneman (1974) showed lead to widespread normative biases:[3]

- *Representativeness*: evaluating probabilities in terms of how representative a particular is of the target category—for example, answering 'How likely is it that $A$ is a $B$?' by considering how similar $A$ is to the prototypical $B$. This can lead people to neglect base rates, sample sizes, and to commit the *conjunction fallacy*. For instance, Tversky & Kahneman (1983) introduced participants to a character, Linda: "Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations." The majority of people judge that it's more probable that "Linda is a bank teller and is active in the feminist movement" than that "Linda is a bank teller." Even though the probability of being a feminist

[2] In Ch. 4, I go on to discuss the possibility of implicit values.

[3] Unlike "implicit biases," it's partly definitive of biases in this sense that they violate the principles of the probability calculus, rational choice theory, or some other presumptively normative psychological rule.

bank teller is necessarily lower than the probability of being any kind of bank teller at all, Linda is more similar to the former category's prototypical member.[4]

- *Availability*: evaluating probabilities according to the ease with which something comes to mind (its "availability"), leading people to overestimate the frequency of events that are salient, dangerous, or highly dramatic, like natural disasters and terrorist attacks. For instance, most people judge that there are more English words beginning with 'K' than words that contain 'K' as the third letter, since the former are more easily brought to mind, even though the latter are in fact 3 times more frequent.

- *Anchoring (and Adjustment)*: evaluating probabilities or other values by starting with some contextually salient number (the "anchor") and then "adjusting" up or down from this anchor to reach an answer—typically, not adjusting far enough. For instance, participants can be given arbitrary anchors between 0 and 100 and then asked to estimate the percentage of African countries that are members of the United Nations by evaluating whether it's lower or higher than the anchor they've been given. The median answer of participants given the initial number of 10 is 25%, whereas participants given the arbitrary anchor of 65 provide a median answer of 45%.[5]

One anchor that's ubiquitous across many contexts is the *status quo*; numerous experiments have shown people implicitly evaluate options (or "alternatives") relative to it. For instance, people take the value of losses (relative to what they currently own, and so the status quo) to be larger than the value of relative gains from this reference point—a phenomenon known as *loss aversion*, and one of Kahneman & Tversky's (1979) primary motivations for developing *prospect theory*—an alternative to rational choice theory (discussed in the next section). Prospect theory departs from rational choice theory precisely in making use of a (context-dependent) reference point.[6] In turn, loss aversion is thought to explain two other heuristics: the *endowment effect* (Camerer 1992, Loewenstein & Issacharoff 1994, Medvec, Madey, and Gilovich 1995) and the *ownership effect* (Beggan 1992), according to which people require more money to sell something they already own than they would be willing to pay to newly acquire it (see also Kahneman, Knetsch, and Thaler 1991).

Loss aversion and the endowment and ownership effects start to show how the implicit *framing* of a decision can drastically influence preferences. The "Asian Disease" problem is among the most dramatic examples of framing effects. Tversky and Kahneman (1981) asked participants to imagine the outbreak of a disease that was expected to kill 600 people, and to choose between two proposals to combat it. Participants were assigned to one of two

[4] "Representativeness" is near-synonymous with similarity or "resemblance," one of Hume's (1738/1978) three principles of association, and so is likely responsible for many of the results discussed in Ch. 2.

[5] As discussed in Section 5, priming participants to anchor on the last two digits of their social security number (SSN) also significantly affects the price they're willing to pay for various goods (Ariely et al. (2003).

[6] The reference point relative to which options count as losses and gains is typically the status quo, but it can be shifted by other implicit factors—e.g., by Ariely et al.'s (2003) SSN anchor.

conditions, both of which involved exactly the same choice, but with the following variation in framing—in this case, the wording of options. When the choice is between:

> If Program A is adopted, 200 people will be saved, and
>
> If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved,

72% of participants choose program A over B. But when the choice is worded as follows:

> If Program C is adopted 400 people will die, and
>
> If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die,

only 22% of participants favor program C over program D, even though A and C (and B and D) are identical, save for how their outcomes are described. Thus, most participants' preferences *reverse* depending on the framing of the decision problem and its options. All manner of framing effects are now copiously-documented and their widespread influence on behavior is uncontroversial. These reversals are one of the most striking instances of where behavioral economics starts to put pressure on rational choice and expected utility theory.

Work on preference reversals actually predates much of the other heuristics and biases research. Lichtenstein and Slovic (1971) showed that participants preferred different options depending on whether they were framed in terms of *betting* on gambles, *choosing* between them, or *rating their attractiveness*, and that by sequentially varying these "methods of elicitation," participants could be induced to make intransitive choices. Because such choices violate the principle of transitivity, these findings seem to undermine a fundamental axiom of rational choice and expected utility theory, insofar as these are understood as predictive theories of behavior. Because actual behavior can easily be made intransitive, it seems that its psychological springs don't conform to rational choice and expected utility theory.

This result was predicted by Slovic and Lichtenstein (1968) and has been repeatedly replicated in a number of subsequent experiments, including in Las Vegas casinos (Lichtenstein and Slovic 1973).[7] Such preference reversals present a serious challenge to rational choice theory and the social sciences that rely on it, especially economics. Some economists have in turn investigated framing effects for themselves, often with the (unintentional) result of gathering further evidence for them. Grether and Plott (1979: 634), for instance, note that "the results we obtained were not those expected when we initiated this study. Our design controlled for all the economic-theoretic explanations of the phenomenon which we could find. The preference-reversal phenomenon which is inconsistent with the traditional statement of preference theory [rational choice theory] remains."

[7]For comprehensive surveys of preference reversals, see Lichtenstein and Slovic (2006) and Hausman (1992: Ch. 13). Hsee et al. (1999) show that reversals can also be induced by asking participants to evaluate options separately vs. jointly, a framing that's been shown to lead to biases in other areas, as well. See also Fischer & Hawkins (1993) and Tversky, Sattath, and Slovic (1988).

## 3.3 Preferences in Rational Choice and Expected Utility Theory

*Rational Choice Theory (RCT)* is *the* main vehicle for explanation and prediction in the social sciences (Elster 2007 and Risjord 2014), especially in its most "scientific" field, economics (Hausman 2012 and Reiss 2013). RCT builds on *expected utility (EU) theory*, which (in its descriptive guise) attempts to explain actions and choices in terms of agents' *preferences*, which these theories take to be completely determined by (i) an agent's subjective beliefs or credences, represented by a probability function, $p$, and (ii) an agent's desires (or strengths thereof), represented by a utility function, $u$.[8] Formally, where $x_1$ and $x_2$ are members of the set $X$ of available alternatives, $x_1 \succ x_2$ *iff* one *strictly prefers* $x_1$ to $x_2$. One *weakly prefers* $x_1 \succeq x_2$ *iff* one strictly prefers $x_1$ to $x_2$ or one is indifferent between these alternatives.

RCT and EU theory traditionally understand preferences as simply "given"—that is, they treat the beliefs and desires represented by the $u$ and $p$ functions as passive, not something an agent reasons or deliberates her way to or as anything capable of further explanatory power at the level of the behavioral and social sciences.[9] As will become clear, this is perhaps the primary flaw with RCT and EU theory from the perspective of behavioral economics.

According to EU theory, for a gamble $g = \{E_1, x_1; E_2, x_2; \ldots; E_n, x_n\}$:[10]

$$EU(g) = \sum_{i=1}^{n} p(E_i)u(x_i)$$

One prefers gambles with higher EUs, and one's preferences obey several axioms:

> *Completeness*: For all $x_1, x_2$ in $X$, either $x_1 \succeq x_2$, or $x_2 \succeq x_1$, or both.
>
> *Transitivity*: For all $x_1, x_2$, and $x_3$ in $X$, if $x_1 \succeq x_2$ and $x_2 \succeq x_3$, then $x_1 \succeq x_3$.[11]

For EU theory and RCT, these axioms are partly definitive of the notion of a preference itself: only by assuming that a person's preferences obey these axioms can they be represented with a probability and utility function at all. If something doesn't obey these axioms, it simply does not count as a preference in the sense recognized by EU theory and RCT.

As Daniel Hausman (2012: 15–16) notes, two additional axioms are in fact needed in order for preferences, so defined, to be capable of explaining and predicting any behavior:

[8] On evaluative and prescriptive versions of EU theory and RCT (Bermúdez 2009), the existence of preferences *qua* genuine psychological attitudes or mental states is more controversial (Dreier 1996). Our concern here is solely with descriptive versions of these theories: those that attempt to explain and predict actual choices and behavior. These are refinements of commonsense folk psychological explanations in terms of belief-desire pairs of the sort familiar from, e.g., Davidson (1963/1980; cf. Pettit 1991 and Hausman 2012).

[9] Facts about beliefs and desires could, for example, still retain relevance for neurobiology.

[10] Following Savage's (1954/1972) decision theoretic framework, $\mathrm{x}_i$ is the outcome yielded by each (mutually exclusive and exhaustive) event $\mathrm{E}_i$ that might result from act $g$ (the option or gamble).

[11] Two additional axioms are often included in particular specifications of EU theory. These axioms are *Reflexivity*: for all $x$ in $X$, $x \succeq x$, and *Continuity*: for all $y$ in $X$, $\{x : x \succeq y\}$ and $\{x : x \preceq y\}$ are closed sets.

> *Choice Determination*: Among the alternatives an agent believes to be available, she will choose an alternative that is at the top of her preference ranking.
>
> *Context Independence*: Whether an agent prefers $x$ to $y$ remains stable across contexts and framings.

According to Choice Determination, choices and behavior (which I'll often abbreviate to "choice behaviors") are completely determined by one's preferences. But if preferences shifted radically across contexts or the ways in which the options were framed, they wouldn't be much help in explaining or predicting action. This is ruled out by Context Independence.[12]

A more specific instance of Context Independence is the *Principle of Independence of Irrelevant Alternatives*: if one prefers $x$ to $y$ in one context (to a given extent), this preference shouldn't change simply because new options are added.[13] An anecdote is often used to demonstrate the principle: Ordering desert at a New York diner, philosopher Sydney Morgenbesser is reportedly informed by the waitress that they have apple pie and blueberry pie. "Apple." A few minutes later, the waitress returns to tell Morgenbesser they also have cherry pie. "In that case," he says, "I'll have the blueberry" (Poundstone 2008: 50). Either switching to cherry or sticking with apple would have made perfect sense, but the addition of cherry shouldn't (rationally) change one's preference between apple and blueberry.

*If* we assume that each of these axioms is obeyed, then RCT holds and we can explain the occurrence of a given action in terms of the agent's (strongly) preferring that action to the alternatives the agent believes to be available. However, behavioral economic and other dual-processing findings seem to show that these assumptions are in fact systematically violated. The results garnering the most attention have been those on preference reversals: because people can be induced to choose $g_1$ over $g_2$, $g_2$ over $g_3$, but then $g_3$ over $g_1$, behavioral economists argue that preference reversals show that people's choice behaviors aren't produced by preferences that conform to Transitivity, in particular. If the attitudes that actually produce behavior violate the Transitivity axiom, but preferences just are defined, in part, as conforming to Transitivity, then it seems that the causes of behavior are not preferences (thereby also undermining the assumption of Choice Determination). This would involve more cases of bypassing and beating—cases in which System 1 attitudes that violate Transitivity work around or overcome one's explicit System 2 preferences or values.[14]

So goes the usual take on how behavioral economics threatens RCT theory and traditional economics, at least. However, recent work suggests that preference reversals are not driven

[12]See also McClennen (1990), who notes that Transitivity alone says nothing about how preferences over different sets of alternatives (and not just those within a single decision problem) must be related.

[13]This principle is more familiar in the context of Kenneth Arrow's (1950, 1951) impossibility theorem in social choice theory, but it's often included in specifications of RCT and EU theory, as well.

[14]Other results apparently violate Completeness, which requires that one always prefer $x$ to $y$, $y$ to $x$, or to be indifferent—i.e., not to switch between these preferences probabilistically. In violation of Completeness, it seems that in some cases people do choose probabilistically—i.e., they have a stable preference that favors $x$ to $y$ in some percent of cases, but $y$ to $x$ in the remaining cases (rather than deterministic preferences that they switch back and forth between in different contexts). See Luce (1959) and Reiss (2013: 42).

by intransitivity of psychological attitudes so much as by framing effects (Hausman 1992). In particular, Tversky, Slovic, and Kahneman (1990) show that much of the phenomenon arises because of "scale compatibility"—the tendency to assign a larger role to values expressed in the same units as questions are asked in. On this interpretation, preference reversals occur not because people have any underlying tendency to violate Transitivity, but because System 1 is prone to framing effects, thereby violating Context Independence.[15] This looks much more like *hijacking*. As Hausman (2012: 114–115) rightly notes:

> [T]he experimental findings... leave decision theorists and economists with a modeling choice. They can continue to require that preference determines choice and to treat factors such as loss aversion, endowment and framing effects, and rationalization as affecting preferences, or they can regard these factors as opening a gap between preferences and choices... I favor the first alternative: to treat preferences as determining choice and to regard the factors that influence choices investigated by psychologists as acting via their influence on preferences.

In effect, the findings can either be interpreted as revealing beating and bypassing, in our terms—which open a "gap" between System 2 preferences and choice behavior (thereby undermining Choice Determination, but potentially preserving the status of preferences as "given"); or we can interpret the findings as revealing widespread hijacking—as cases in which System 2 preferences produce behavior (preserving Choice Determination), but where these preferences are in turn influenced by implicit System 1 attitudes that are subject to framing effects and other heuristics. Like Hausman, I favor the hijacking alternative. At the very least, RCT and EU theory only retain any real descriptive power on this interpretation.[16] If preferences of the sort characterized by RCT and EU theory aren't what actually produce choices and behavior, then these theories lose any explanatory and predictive interest. On the hijacking interpretation, however, we can no longer treat preferences as "given"—RCT and EU theory must be supplemented with an account of *preference formation*.

To the extent that RCT and EU theorists address how preferences are formed, they tend to assume it's an entirely rational, reasons-responsive process (Dietrich & List 2013a, 2013b). But the behavioral economic results suggest that this assumption is violated. More importantly for our purposes, for preferences to support *autonomous* (not just rational) behavior, preference formation should occur in accordance with one's values. But the behavioral economic results, I argue, show that our preferences are often formed and influenced not only by arational factors, but in context-dependent ways that one does not value, take there to be reason for, or take to be justified in any way. The real threat is not that people have

[15] Indeed, in his earlier career as a psychologist-cum-behavioral-economist, Donald Davidson pioneered some of the empirical work on preference reversals (see Davidson, McKinsey, and Suppes 1955, Davidson & Suppes 1957, and Davidson & Marschak 1959). In his later philosophical work, Davidson (1974/1980) suggests precisely that preference reversals may not threaten Transitivity so much as Context Independence.

[16] The same might be said for ordinary folk psychology, which takes beliefs and desires to be subject to rational constraints and to *rationalize* (make rational sense of) choice behavior.

preferences that violate the axioms of RCT and EU theory, but instead that the formation of preferences is often hijacked out of the control of what one values and cares about by implicit System 1 attitudes. Without understanding the heuristics or associative guidelines according to which these implicit attitudes go about constructing preferences, we cannot explain the deviations in behavior from normative rules observed in many behavioral economic results.

Part of the reason this deeper threat to autonomy has been overlooked may be the tendency to conflate preferences with valuing and caring about. Valuing and caring are typically attitudes that are stable over time, and are moral central constituents of the practical standpoint of the agent or her self (Bratman 2007, 2009a, 2009b). Preferences are often thought to have these characteristics, as well. Ullman-Margalit & Morgenbesser (1977) claim that "picking" (e.g., selecting one can of Campbell's soup off the shelf rather than any other) can't be conflated with "choosing" because only choosing involves making a selection based on one's preferences.[17] What the behavioral economic results suggest, however, is that if preferences determine actual choice behavior, then preferences must be much more mercurial things—attitudes that may be affected by valuing (and reflection), but that are also widely affected by System 1 framings whose influence we're often unaware of and would not endorse if we were. Of course, we can call valuing and caring "preferences" if we wish. The important point is that we should distinguish what one actually values and sees as good (which I assume is more a dispositional matter) from one's current conception of THE GOOD (the proximal cause of one's behavior). I'll continue to reserve "preference" for the latter. By whatever name, I take it we sometimes recognize (later) that what we thought we valued at the time of action was not what's actually valuable by our own lights.[18]

## 3.4 The Framing and Construction of Preference

The growing consensus in the behavioral and social sciences is that preferences are not "given" at all, but instead "constructed," often on the fly and in ways that are highly sensitive to context-dependent framing effects (Hausman 2012: 111, Lichtenstein and Slovic 2006). As already mentioned, this is now widely thought to explain the phenomenon of preference reversals: if people's preferences or conceptions of THE GOOD are often constructed on the fly, rather than from longstanding, stable dispositions, it's unmysterious why they're sometimes constructed such that $g_1 \succ g_2$ and $g_2 \succ g_3$, but at other times such that $g_3 \succ g_1$. Other work supporting the "construction of preference" is abundant.

Another classic of the behavioral economics literature is Shafir, Simonson, and Tversky's (1993) work on "reason-based choice," which attempts to explain the formation of preferences in terms of agents' consideration of reasons. Shafir et al. (1993) are primarily concerned to contrast a "reason-based" approach with EU theory and other explanatory frameworks which understand preferences simply as "given." Their term "reason-based choice," however, belies the fact that many of their results demonstrate serious distortions in preference-formation

[17] For critical discussion, see Murray & Buchak (under review).

[18] In Ch. 4, I argue that select implicit attitudes actually count as implicit forms of valuing and caring.

and decision-making—deviations from reasons or rational rules that arise precisely when the agent thinks about what reasons she has to justify her choice. In other words, the body of work Shafir et al. focus on suggests that even when one is thinking about what reasons one has—indeed, sometimes precisely *because* one is thinking about one's reasons or what's rational—one's practical deliberation and subsequent choice can be hijacked out of the control of one's valuings and what one genuinely countenances as reasons.

Preference formation and choice are particularly subject to the attributes (or properties or dimensions) along which options differ. In most choice sets, the options differ along more than one attribute—price but also performance and durability, for instance. In order to calculate the total expected utility of each option, an agent must determine how to weight the relative importance of these different attributes against one another. As we already saw in Chapter 2, the process of determining the trade-offs between attributes is an especially strategic place for System 1 attitudes to hijack System 2. Because of the difficulty and frequent ambiguity of how precisely to *weight* each attribute, let alone the *comparative* importance of each attribute, System 1 can interfere here with relatively less chance of detection than by tampering with more transparent reasoning processes (deductive inferences, perhaps).[19]

One set of results that Eldar Shafir et al. (1993) focus on is Paul Slovic's (1975, 1990) research showing that, when faced with options of equal expected utility (i.e., equal EU once each option's weighted values on all attributes have been summed), people often make their choice on the basis of whichever attribute seems most (contextually) important. Shafir et al. suggest this is because an option's being superior on the most important attribute seems like a reason to choose it. This may be a reasonable-enough tie-breaker, but the heuristic can also lead to biases when the options *don't* have equal EU. Shafir (1993) shows that options which provide both compelling pro- and con- reasons can be both chosen and rejected more often than less tantalizing options, depending on which attribute is contextually salient, and Tversky & Shafir (1992a) show that people's desire to find reasons for their decisions can lead them to pay for information that won't actually affect their decisions.[20] More surprising is work on the *asymmetric dominance (or decoy) effect*: a preference for $x$ over $y$ can be increased (made more extreme) by adding a third alternative, $z$, which is clearly inferior to (dominated by) $x$ but not by $y$.[21] These studies show that people sometimes violate the Principle of Context Independence (as well as the Principle of Irrelevant Alternatives).

For instance, Tversky & Shafir (1992b) show that given the choice between \$1.50 (the default) and a Zebra pen, most participants choose the pen. But when given the choice

[19]That said, overall assessments of options can also rationally inform assessments of specific attributes' comparative importance, especially in tie-breaking cases. For instance, candidate $X$ might be superior to candidate $Y$ on dimension $p$, whereas $Y$ is superior to $X$ on dimension $q$. Suppose one has no independent idea whether attribute $p$ or $q$ is the more important skill for a certain job. One might nonetheless know that candidate $Y$ is overall more skilled at that job (for instance, based on $X$ and Y's past performance), and that might give one reason to conclude that the attribute that candidate is superior on, $q$, is in fact more important for that job. Thanks to Tania Lombrozo for bringing cases of this sort to my attention.

[20]This happens when there are "disjunctive reasons." See also Shafir & Tversky (1992).

[21]This was first demonstrated by Huber, Payne, and Puto (1982). See also Wedell (1991).

between $1.50, the Zebra pen, and two Pilot pens, only about half choose *either* of the pen options. Most choose the $1.50 (even though in both conditions participants are told that all pen options have a value just over $2.00). Similarly, Simonson & Tversky (1992) show that the number of participants who choose a Cross pen over $6 increases when a considerably less attractive pen is added as a third option. Asymmetric dominance is also observable in many "natural experiments": When Williams-Sonoma added a second electronic bread-maker to their catalogue, priced at $249, they didn't sell many; but this move did (unwittingly, in their case) nearly double sales of the $279 bread-maker that they'd offered all along.[22]

Asymmetric dominance also occurs in dating (and mate) selection: Dan Ariely (2008: 14) finds that 75% of people report that they would rather go on a date with a person depicted in a picture A rather than the person in picture B when they're also given a third option, a "decoy" of A that's been photoshopped to look slightly less attractive. Ariely recommends finding a "wingman" who looks similar to but slightly less attractive than you. Sedikides, Ariely, and Olsen (1999) find that in choices about who participants would like to date, the effect is driven by pre-decisional biasing of the importance of different attributes.[23]

Shafir et al. (1993) suggest that the asymmetric dominance effect is explained by participants' attempt to find a seemingly compelling reason or justification for their decisions: the presence of $z$ is normatively irrelevant to any preference between $x$ and $y$, but $z$'s clear inferiority to $x$ may, at the level of implicit heuristics, seem to provide a reason in favor of $x$. Another set of results show that features of an option that an agent cannot justify may weaken any preference for that option. Simonson, Nowlis, and Simonson (1993) show that experimental participants are less likely to choose an alternative that they know was chosen by another person whose reason for choosing it doesn't apply to themselves (e.g., proximity to one's family) compared to when participants are not told why the other person made their choice. Simonson, Carmon, and O'Curry (1994) show that adding positive but unneeded or unwanted features to an option can also decrease preferences for that option, even though participants know that the "bonus" is in effect costless. So much for promotional offers.[24]

Shafir, Simonson, and Tversky (1993) often appear ambivalent about whether "reason-based choice" involves what I labeled "advising" in Chapter 2 or whether it instead demonstrates the existence of hijacking (or other forms of conflict like beating and bypassing). For instance, they note that when the axioms of RCT and EU theory transparently apply to a situation, people overwhelmingly tend to choose and behave in accordance with those axioms (Tversky & Kahneman 1986, Tversky & Shafir 1992a). But, of course, often these decision

[22]This effect is reported in Shafir et al. (1993: 25) and Ariely (2008: 14).

[23]See also Ariely & Wallsten (1995), who demonstrate that participants both distort the values of attributes themselves and distort the attributes' comparative importance or weighting. For earlier work showing that the search for dominance is a decision-making heuristic more generally, see Montgomery (1983, 1989, 1993) and Montgomery & Svenson (1983). Conversely, people also display *extremeness aversion*: options with extreme values are perceived as being less attractive than those with more intermediate values, presumably because the intermediate values' status as a "compromise" between the extremes provides an apparent reason for choice (Simonson 1989, Simonson & Tversky 1992, and Tversky & Simonson 1993).

[24]Shampanier, Mazar, and Ariely (2007) show that people can be induced to make the converse mistake in other contexts, as well: making suboptimal choices in favor of free options.

rules are not transparent, and Shafir et al. (1993: 33) do readily add that "[r]easons, it appears, lend themselves to certain framing manipulations." Indeed, Shafir et al. (1993: 34) conclude by distinguishing the types of *pre-hoc* rationalization and distortion revealed by the results they draw on from the more familiar *post-hoc* variety (as I did in Chapter 2):

> These results suggest that the axioms of rational choice act as compelling arguments, or reasons, for making a particular decision when their applicability has been detected, not as universal laws... it appears that people often do not have well-established values, and that preferences are actually constructed—not merely revealed—during their elicitation (cf. Payne, Bettman, & Johnson, 1992). A reason-based approach lends itself well to such a constructive interpretation. Decisions, according to this analysis, are often reached by focusing on reasons that justify the selection of one option over another. Different frames, contexts, and elicitation procedures highlight different aspects of the options and bring forth different reasons and considerations that influence decision.

The classic dual-processing picture that might initially look most promising for RCT and EU theory, then—"*reason*-based choice"—actually turns out to be anything but. On Shafir et al.'s (1993) picture, "reasons" are constructed on the fly and often hijacked by features of whatever situation one finds oneself in that one would not take to provide any justificatory force. Further work confirms that implicit attitudes can bias the processing of attributes.

For one, focusing on particular attributes can bias the information and so *belief* component of preferences in a way that influences choice behavior. J. Edward Russo and colleagues (1996) induced some participants to prefer one alternative over another, and then gave them information about various attributes of those alternatives, one at a time. Participants were next asked to rate the extent to which each attribute favored one alternative over the other. What they found was that participants were more likely to say the attributes favored their preferred alternative than participants in a control condition. More surprisingly, Russo, Meloy, and Medvec (1998: Exp. 1) found that such predecisional distortion of information doesn't require any *antecedent* preference. In a condition in which there was no such preference, participants were asked which alternative was currently "in the lead" (which alternative they would choose if they were, hypothetically, forced to choose at that moment) at the same time they were asked to rate the extent to which each attribute favored each alternative. Even in this condition, participants tended to answer that attributes favored whichever alternative was currently in the lead. In other words, people distort their interpretation of information in such a way that it favors an *emerging* preference, even when there is no antecedent preference. Moreover, participants seem to distort information *more* when it's more balanced and doesn't clearly support either alternative (Russo et al. 1998: Exp. 2, 2000). Chaxel, Russo, and Kerimi (2013) show that people's search for information about alternatives is also biased in favor of the option currently in the lead. Together, these

studies strongly support the hypothesis that people distort information so as to *rationalize* preferences that they have not even formed yet—preferences emerging from System 1.[25]

These sorts of motivated theoretical reasoning, even in the context of practical deliberation, are one thing. More scandalous are results showing that people's desires and very conception of THE GOOD are "constructed." Simon, Krawczyk, and Holyoak (2004) present some of the best known findings. They gave participants a (pre-test) series of questions about different attributes of hypothetical job offers, asking them to rate the desirability of jobs with various attributes and the attributes themselves. Participants were then told about two specific jobs, which differed along four attributes—commute time, whether or not the job would come with a private office or only a cubicle in a shared office, salary, and amount of vacation time. Finally, participants were asked which job they would decide to take, and another (post-test) series of questions about the specific attributes of those jobs.

Simon et al. (2004) found that participants' judgments of overall goodness (the total EU computed from their post-test questions about the importance of different attributes) predicted their decisions about which job to take well, and that participants were confident about these decisions. However, participants' assessments of overall goodness and their decisions were not determined by set, pre-test preferences. Instead, participants latched on to one attribute as particularly salient early in the whole process, and then adjusted the relative importance of the other attributes (and their desirability) in whatever way would justify that initial inclination, as demonstrated by pre-to-post-test changes.

Specifically, Simon et al. find evidence that the desirability of particular attributes and the desirability of jobs with those attributes is strengthened for those that support a participant's emerging preference, and that the desirability and importance placed on the attributes favoring the less desirable job are weakened during the process of deliberation. In other words, participants rationalize their preferences for options *while forming them* by shifting their judgments about the desirability and importance of those options' (and the competing options') attributes in whatever way would justify choosing in favor of their still-emerging preference.[26] This is clear evidence of hijacking: participants' preferences are being formed in ways that they presumably do not value and would not take to be justified. One's System 2 preferences and what one comes to see as GOOD may causally determine what one chooses to do and does, but when these preferences were formed, they were hijacked from the control of what one valued and cared about. In these cases, preferences are driven by implicit System 1 inclinations, which turn one's processes of practical reasoning into pre-hoc rationalizations for their own "ends," bending deliberation and evaluation in sophisticated ways to make the decision one's System 1 is still in the process of reaching seem justified. Once formed, these preferences are then available as inputs to subsequent post-hoc rationalizations, as well.

Summarizing their results, Simon et al. (2004: 335) write:

[25]See also Brownstein (2003) and the associated references in Ch. 2.

[26]Compare these findings to those showing that implicit biases can hijack judgments about the importance of different criteria for college admission discussed in Ch. 2 (Dovidio and Gaertner 2000, Hodson, Dovidio, and Gaertner 2002, Uhlmann and Cohen 2005, and Son Hing et al. 2008).

> In general, both the reported values of the attributes (ratings of desirability) and their weights (ratings of importance) shifted to make one alternative dominate the others... These findings cannot be attributed to differences in methods used to elicit or describe the options, nor to variations in context (cf. Slovic, 1995; Tversky & Kahneman, 1986). Rather, the reconstruction of preferences seems to be the natural outcome of the very process of decision making.

In other words, contextual "framing" by the external situation is only the tip of the iceberg—by their very nature, most processes of preference formation are affected by implicit System 1 attitudes, opening the door to hijacking and other forms of conflict of the type catalogued in Chapter 2. As Hausman (2012: 111–2) puts the upshot of Simon et al.'s (2004) and similar findings: "subjects adjust their preferences to rationalize their choice, which in turn results mainly from an evaluation in terms of a single dimension that the experimental subject finds particularly salient," where salience is contextually mercurial and influenced by heuristics.

Consider one final, particularly striking example. Carlson, Meloy, and Russo (2006) show that the *order* that attributes are presented in can bias people's choices.[27] Russo, Carlson, and Meloy (2006) extend this finding, showing that if attributes are presented in the right order, participants can be induced to choose the alternative that they actually value *less* overall. Participants were first presented with a series of restaurants that differed along 9 attributes (e.g., atmosphere, dishes served, distance away) and asked to make pair-wise choices between them, establishing participants' actual comparative evaluations and also giving the researchers information about which attributes participants valued most.

Russo et al. then renamed the restaurants and recombined the information in their descriptions into 6 attributes in such a way that—based on their work on predecisional belief distortion, discussed above—they predicted would bias people in favor of what had been, by their own earlier pair-wise lights, the inferior alternative. Specifically, they presented the attribute that favored the inferior option first, in order to establish it as the leader, and presented the next most favorable attribute for this option last, to also give it the advantage of any recency effect. Participants returned two weeks later and were given the renamed, modified materials. As they were presented with each attribute, participants were asked how strongly it favored either restaurant, and then asked to choose which restaurant they'd rather dine at. Russo et al. found that the attribute order effect biased participants' assessment so much that over half of them ended up choosing the restaurant that, by their own (earlier) lights, they preferred less—that is, that they chose the other restaurant instead of before.

How are we to describe such a case of preference reversal? "Bypassing" would suggest that participants are somehow making their choices in a way that circumvents their reasoning and explicit evaluation of the (relative) goodness of the options entirely. But given the nature of the task, it's clear this isn't what's happening: participants are explicitly engaged in just this type of (highly reflective) practical reasoning and deliberation, and there's no reason to think it's epiphenomenal. Instead, we should see these results as further evidence of hijacking:

[27] Pennington & Hastie (1988) demonstrate a similar order effect in the legal context.

participants' preferences are precisely what determine their choices, but given the subtle knowledge Russo et al. (2006) have about the way preferences are formed, they're able to present the alternatives in such a way that allows System 1 to wrest people's preferences and reasoning out of the hands of what they actually care about and value, at least comparatively (in this case, the stronger value they assign to eating at the other restaurant).

Similar shifts in the weighting of different attributes may drive preference reversals more generally (Fischer & Hawkins 1993, Hawkins 1994, Tversky, Sattath, and Slovic 1998). As mentioned above, people are implicitly influenced by "scale compatibility." They tend to assign a larger role to values expressed in the same units as a decision problem is presented in. "Noncompatibility" between the units that the input and output are described in requires additional calculation, and so increased effort that System 2 shies away from (Kahneman 2011). Thus, attributes expressed in the same (or more similar) units to the question asked (or the output required) are weighted more heavily. Depending on how a choice is framed—what units it's described in—one or another attribute might be weighed more heavily in one's practical reasoning because of the units it's described in, producing different preferences.

There are surely cases in which factors other than one's preferences produce choice behavior—cases of beating and bypassing. But the results discussed in this section suggest that even when Choice Determination and the other axioms of RCT and EU theory hold, Context Independence often does not: many preferences are merely constructed as *pre-hoc* rationalizations of whatever System 1 finds most salient or representative. The risk of hijacking is especially high in cases where options have to be compared along different dimensions or attributes. Because of the cognitive difficulty and complexity of weighting and trading these attributes off against one another, System 2 is less likely to catch System 1 in the act of hijacking in these compared to dimensionally simpler choices.

## 3.5 Coherent Arbitrariness

It may seem that we face a choice between RCT and EU theory's conception of preferences as contextually stable, "given," pre-existing entities (relative to the time of deliberation and decision), on the one hand; and, on the other, the behavioral economist's conception of preferences as completely transient attitudes constructed "on the fly" by heuristics, framing, othe variable features of context, and the processes of deliberation. Plausibly, though, the truth lies somewhere in between, as Dan Ariely and colleagues have argued. In a series of experiments, Ariely et al. demonstrate some of the most striking instances of anchoring and arbitrary preference construction, but in the process they show that these initial "imprintings" have stable and sensible downstream effects, a thesis they call *coherent arbitrariness*.

Ariely, Loewenstein, and Prelec (2003) asked MBA students at MIT's Sloan School of Management to write down the last two digits of their social security numbers (SSNs) as a dollar amount next to each of six products (a cordless trackball, cordless keyboard and mouse, an average bottle of wine, a rare bottle of wine, Belgian chocolates, and a design book), and then asked them whether they'd hypothetically be willing to pay that price for each product

("yes" or "no"). Participants next specified how much they would actually be willing to pay, or "bid" for each product, and the highest bidder then in fact did pay that much for the item and subsequently received it. Even though participants were explicitly reminded that their SSNs were entirely "random" and conveyed no important information, participants with higher SSNs were willing to pay significantly more than those with lower SSNs—that is, their SSN acted as an anchor. Participants with SSNs ending 80-99 placed "bids" that were 216–346% higher than those with SSNs ending 00-19. For example, participants in the former group were willing to pay $56 on average for the cordless keyboard and mouse; those with lower SSNs only an average of $16. However, all participants made coherent "bids" within categories—e.g., they were willing to pay more for the keyboard and mouse than the trackball, and more for the rare bottle of wine than the average bottle. These are dramatic effects, and as Ariely (2008: 28) notes: "[s]ocial security numbers were the anchor in this experiment only because we requested them. We could have just as well asked for the current temperature or the manufacturer's suggested retail price (MSRP)."[28]

In another experiment, Ariely et al. (2003) had participants listen to 30 seconds of an annoying sound (e.g., a 3,000hz sound approximating "a high pitched scream") and then asked them whether they'd hypothetically be willing to listen to that sound again for either 10 cents or 90 cents, where these monetary amounts acted as the initial anchors. Next, participants indicated how much they would have to be paid to listen to a similarly annoying sound, which they did then listen to and were paid for. Both groups then completed another trial in which the anchor was 50 cents, and then a third trial where the anchor was the opposite of what they'd first encountered (e.g., participants in the group given an anchor of 10 cents on the first trial were asked in the third trial whether they'd be willing to listen to the sound again for 90 cents, and vice-versa). The goal of this experiment was to see whether people would switch or adjust their anchor to each new context and choice, in effect updating their anchor, as more extreme construction of preference views might predict, or whether people's decisions would instead be more influenced by, and cohere with, the initial anchor from the first trial, in line with the thesis of coherent arbitrariness.

What Ariely et al. found was clear support for the latter. Ratings on the first trial were influenced by the initial anchor (with participants in the 10 cents condition requiring 33.5 cents on average to listen to the sound and those in the 90 cents condition demanding 72.8 cents on average). However, answers in the third trial were also closer to the first trial anchor than they were to the last anchor, with participants initially given the 10 cents anchor requiring 45.3 cents on average to listen to the annoying sound in the third trial and those initially given the 90 cents anchor demanding an average of 63.1 cents.[29] Thus, even for those exposed to the same three anchors, the initial anchor still had a significant effect on decisions, three trials later, and a stronger effect than the last anchor. This is strong support for coherent arbitrariness: people adjust or update their preferences in a sensible, seemingly

[28]It's probably an overstatement to say that *any* numbers could serve as (equally powerful) anchors, since most individuals tend to have "implicit self-associations" (discussed in the next chapter) with their own SSN.

[29]In the second trial, participants in the initial 10 cents condition demanded an average of 43.5 cents and those in the initial 90 cents condition demanded an average of 63.2 cents.

rational way to cohere with their past choices and initial preferences, knowledge of which they use as inputs to current decision-making (Gilboa & Schmeidler 1995). But those initial preferences are often the arbitrary product of whatever framing effects and other contextual features happened to be present when one first encountered the relevant stimulus.

As Ariely et al. (2003: 74-5) put it: "valuations are initially malleable but become 'imprinted' (i.e., precisely defined and largely invariant) after the individual is called upon to make an initial decision... Following imprinting, valuations become locally coherent, as the consumer attempts to reconcile future decisions of a 'similar kind' with the initial one. This creates an illusion of order, because consumers' coherent responses to subsequent changes in conditions disguise the arbitrary nature of the initial, foundational, choice" (see also Hoeffler & Ariely 1999). People scale subsequent preferences appropriately to similar initial preferences, but ultimately, initial preferences are highly subject to non-normative factors.

Ariely, Loewenstein, and Prelec (2006) take this research one step further. Most work on anchoring and framing effects shows that the price participants are willing to pay for an item—that is, *how* good they currently conceive the item to be—can be dragged about by normatively arbitrary factors. Ariely et al. (2006) ask a more basic question: can *whether* one sees something as good vs. bad in the first place—whether they're willing to pay for it, rather than having to be paid for it—similarly be pushed around by normatively arbitrary factors? This work is inspired by *The Adventures of Tom Sawyer*, in which Tom convinces his friends that they want to help him whitewash a fence because he puts on considerable airs of enjoying the task himself.[30] Ariely et al. (2006) ask "*Do people even have a pre-existing sense of whether an experience is good or bad*?" or, following "Tom's Law," is whether one sees something as valuable or disvaluable itself even subject to coherent arbitrariness?

In the first experiment, participants were told that Dr. Ariely would be giving a 15 minute poetry recital from Walt Whitman's *Leaves of Grass* in a week's time. Half the participants were asked whether they'd hypothetically be willing to pay $2 to attend the poetry recital. The other half of participants were asked whether they'd hypothetically be willing to attend

[30] "Oh come, now, you don't mean to let on that you *like* it?"

The brush continued to move.

"Like it? Well, I don't see why I oughtn't to like it. Does a boy get a chance to whitewash a fence every day?"

That put the thing in a new light...

"Say, Tom, let *me* whitewash a little."

...There was no lack of material; boys happened along every little while; they came to jeer, but remained to whitewash. By the time Ben was fagged out, Tom had traded the next chance to Billy Fisher for a kite, in good repair; and when *he* played out, Johnny Miller bought in for a dead rat and a string to swing it with—and so on, and so on, hour after hour. And when the middle of the afternoon came, from being a poor poverty-stricken boy in the morning, Tom was literally rolling in wealth...

He had discovered a great law of human action...There are wealthy gentlemen in England who drive four-horse passenger-coaches twenty or thirty miles on a daily line, in the summer, because the privilege costs them considerable money; but were they offered wages for the service, that would turn it into work and then they would resign (Twain 1876: Ch. 2).

the recital if they *received* $2 compensation. These were the anchors. Participants were later told that the recital would in fact be free, and were asked if they'd like to receive an email about its time and location (to determine whether they saw the experience as positive—something they might like to attend). 35% of participants in the condition that suggested the experience was good or valuable (something to potentially pay for) wanted to receive the email, compared to only 8% of participants in the condition whose anchor suggested that the experience was bad or disvaluable (something you'd need compensation for doing). As with Tom's friends, it seems, the mere suggestion that something is good (or bad) is often enough to frame whether it falls under people's conception of THE GOOD (or THE BAD).

A further experiment tested whether these initial arbitrarily-influenced preferences would lead to future preferences that "cohered" with the initial anchors. Half of the participants were asked whether they would hypothetically be willing to pay $10 to hear Ariely recite poetry for 10 minutes, while the other half of participants were asked whether they would hypothetically be willing to hear Ariely recite poetry for 10 minutes if they received $10 compensation. All participants were then asked how much they would be willing to pay (or be paid, respectively) for 1, 3, and 6 minutes of the poetry recital. Participants who were asked how much they would pay were, on average, willing to pay for the experience, whereas participants who were asked how much they would need to be paid on average required compensation to undergo the experience, replicating the findings of the first experiment.

However, participants in both conditions of the second experiment consistently indicated higher sums of money for longer durations; i.e., they required being paid more money, or were willing to pay more, for 6 minutes of the experience than for 3 minutes, and more for 3 minutes than 1 minute, showing that despite the initial arbitrariness of whether they saw the recital as good vs. bad, participants scaled subsequent preferences to cohere with this initial preference.[31] A third experiment replicated these results, despite the fact that participants first heard a 1 minute sample of the recital (which should decrease the effect of the anchor), and were put into either the "willing to pay" or "willing to be paid" conditions in a way that was transparently arbitrary (obviously based on their SSN), *and* despite each participant's answering both sets of questions—such that they knew what participants in the other condition were asked to do (which should eliminate any inference that the initial question provides information about the actual pleasurability of the experience).[32]

Ariely sometimes makes the effects in these experiments sound like those of bypassing, speaking as if our initial decisions, at least, are not the product of our preferences at all.

[31] Ariely et al. (2003) demonstrate a similar finding in the sound experiments: participants were first asked how much they would require to listen to 100 seconds of a sound, next asked about 300 seconds, and third 600, or vice-versa (asked the 600 second question first, then the 300 seconds, and the 100 second question last). The group starting with 100 seconds required an average of $3.78, $5.56, and $7.15 to undergo the respective noise durations, whereas the group starting with 600 seconds required an average of $5.16, $3.65, and $2.01, showing how coherent arbitrariness can be compounded over time.

[32] Ariely (2008: 42) says "[t]he students did not know whether listening to me recite poetry was a good or bad experience, but whatever their first decision was, they used it as input for their subsequent decisions and provided a coherent pattern of responses across the three poetry readings."

Ariely admits, though, that this makes it mysterious just what motivational states *do* produce behavior in these cases. Again, I think, it's better to see this as hijacking: cases in which one's preferences are precisely engaged, but in which one's initial preferences, at least, are not formed in a way that's dependent on one's valuings or reasons (but instead on various System 1 attitudes one would not endorse). These initial choices can then have self-reinforcing effects, which may be coherent. Indeed, Hoeffler and Ariely (1999) show that the more one chooses a particular option, the stronger one's preference for it becomes. Their data also suggest that preferences stabilize as the result of having to make repeated difficult trade-offs because this leads to a more coherent conception of the relative importance of different attributes. In this way, Ariely (2008: 36) notes, "our first decisions resonate over a long sequence of decisions." "The power of the first decision can have such a long-lasting effect that it will percolate into our future decisions for years to come" (Ariely 2008: 44; cf. Ariely & Norton 2008). Even if preferences are adjusted or updated in a perfectly rational, coherent way, initial violations of Context Independence can be perpetuated throughout numerous subsequent choices.

Of course, poetry recitals may be small stakes, and some experiences are manifestly good or bad. But many of our most important decisions—about who to date or who to marry, what career to pursue or which state to move to—require drawing on a number of different experiences that are highly ambiguous and subject to anchoring and other framing effects, especially when multiple attributes must be compared at once (Ariely & Carmon 2003). We may worry that even our most stable, central preferences ultimately trace back to initial factors other than what we value and care about—factors we might not take to be normatively relevant in the least, let alone to constitute sufficient justification.[33] The relevance of these findings extends to financial markets—the initial price of stocks is often ambiguous and subject to similar framing, like conformity effects (though later adjustment of stock prices may be coherent), and to labor markets—employees are often highly aware of salary adjustments in their own workplace, but have no sense of what their labor is worth in absolute terms or relative to what they could make at other firms (Ariely et al. 2003).

In a way, then, the threat from behavioral economics is less dire than it first appears: it's not so much that our preferences are always beaten or bypassed, or that they fail to obey the axioms of RCT and EU theory, or that they're inherently unstable. They may sometimes be bypassed, and at least initially, when first encountering options of a given sort, they may not be stable. Ariely et al.'s work suggests that after repeated experience with options, though—especially after repeatedly choosing the same options and having to make trade-offs between different attributes (Hoeffler & Ariely 1999)—stable preferences may develop.[34] However, in another way, the threat from behavioral economics is even worse: we may eventually develop stable preferences, but often these only carry over and compound the arbitrary framing of whatever contexts we first encountered those or similar stimuli in.

[33]Presumably, most participants not only explicitly judge that their SSN does not provide any reason to pay or be paid any particular amount for anything; they also do not take their SSN to be normatively relevant whatsoever (on any—including implicit—level, nor from within any fragment or subset of their mental states). Indeed, participants deny that their SSN has any effect on their choices (Ariely et al. 2003).

[34]See also Hammond et al. (1975, 1980) and West (1996).

## 3.6 Conclusion

Only further research can determine the precise extent of these influences, but the findings presented here are far from atypical. "There are probably hundreds of articles that could be cited as illuminating the nature of preference construction" (Lichtenstein and Slovic 2006: 24). Even if the proponent of RCT and EU theory can explain away some findings, the general picture coming out of behavioral economics (like social psychology) is clear: the ultimate sources of human choice behavior often ground out in something other than what we value. Even when there are preferences in play, these often begin as *pre-hoc* rationalizations, which then become *post-hoc* rationalizations as future preferences are adjusted to cohere with them. This replicates the conclusion (of Chapter 2) that hijacking poses a threat to autonomy, but it goes on to show just how deep and persistent that threat can be.

Not only is explicit System 2 *reasoning* shot through with the influence of System 1 heuristics and biases, but even one's current, working conception of THE GOOD can be framed, and thus hijacked from the control of what one values and cares about. Moreover, these initial hijackings can persist and grow in influence. Even if one's System 2 reasoning takes over in an uncompromised way forever after—and is solely responsible for the construction of any new, similar preferences in a way that fully accords with the axioms of RCT and EU theory—so long as later reasoning and preferences are made to cohere with initial preferences whose formation was arbitrary, the arbitrariness is only further compounded.

Many frames and other heuristics revealed by behavioral economics are easily harnessable by agents bent on intentional manipulation. As Ariely (2008: 45) notes, "[i]n the real world, anchoring comes from manufacturer's suggested retail prices (MSRPs), advertised prices, promotions, product introductions, etc."[35] And while some of these influences may even have started unintentionally, advertisers and others are of course watching the work coming out of behavioral economics with great interest. There is growing recognition that manipulation occurs as a matter of course in the actual economic world (Akerlof and Shiller 2015) and behavioral economics may only lend advertisers and others more effective strategies.[36]

The behavioral economic results also raise a classic philosophical issue about whether there are historical conditions on autonomy anew (Fischer & Ravizza 1998, Watson 2001), but in a way that's also empirically rich. Granting that intentional manipulation and implicit influences can hijack our reasoning and preferences out of the control of our valuings, can what we value and care about itself be hijacked? Consider Ariely's (2008: 43) worry:

> Could it be that the lives we have so carefully crafted are largely just a product of arbitrary coherence? Could it be that we made arbitrary decisions at some point in the past... and have built our lives on them ever since, assuming that the original decisions were wise? Is that how we choose our careers, our spouses,

[35] According to Ariely (2008: 30), however, one also has to make a choice or choices that conform with these pricing "suggestions" before they can become actual anchors that affect further preferences.

[36] See also Murray & Lombrozo (2017) and Shafir (2016). These issues are also relevant to whether the government should "nudge" people using dual-processing research, discussed further in Ch. 6.

> the clothes we wear, and the way we style our hair? Were they smart decisions in the first place? Or were they partially random first imprints that have run wild?... suppose we are nothing more than the sum of our first, naive, random behaviors. What then?

What then, indeed. The initial questions are unsettling enough. But consider the last. Suppose one has a preference that was formed arbitrarily—that conflicted with one's values at the time of its formation—but which has, perhaps over the course of many years, gradually brought many of one's values themselves into coherence with it in its wake. Would acting on that preference now, given that it conforms with one's values (and has shaped many of them), still threaten one's autonomy, given that preference's arbitrary origins? Structuralist or *ahistorical* accounts of autonomy (which count only facts about an agent's *current* psychological economy as being relevant to autonomy) would say "no" (Frankfurt 1971/1988, Watson 2001). According to ahistorical accounts, even if one's preferences or valuings were initially the product of framing or other untoward factors, one is nonetheless currently responsible for acting on them, if and insofar as they've genuinely become one's values and integrated themselves into what one cares about. I don't aim to take a stand on the debate between historical and ahistorical accounts of autonomy here, only to note how the issues discussed in this chapter can inform and be informed by traditional debates about agency. In the next chapter, I turn to whether there can be implicit values and forms of caring.

# Chapter 4

# Identification by Association

## 4.1 Introduction

As I write, cases of the following kind dominate U.S. news:

> A young black man, call him $Z$, and police officer $X$, who is white, have a split-second altercation in which $X$ shoots and kills $Z$.

2014 is the summer of Michael Brown, Eric Garner, Tamir Rice, and the nationwide protests following their deaths. The summer of 2017 is much the same.

The problem may seem obvious, but all of its causes are not. Levels of explicit racism in the United States have only dropped precipitously since the 1960s, yet marked racial disparities in the use of violent police force and numerous other forms of discrimination persist, crying out for explanation. Much of this explanation has been recently unearthed in the guise of implicit bias. As discussed in previous chapters, an entire literature now warns that implicit attitudes influence a frightening array of human behavior. Among the most sinister influences: implicit racial biases are implicated in cases like those above, raising the likelihood of mistaking tools for guns and shooting those holding them (Payne 2001).

Implicit biases operate below the radar of conscious awareness and often resist one's explicit judgments and any attempt to change them. On the other hand, there's clearly a moral problem here and, some say, culprits. The following has become a pressing question:

> If $X$ shoots $Z$ because of an implicit racial bias, is $X$ blameworthy for doing so?

More generally, is one morally responsible for acting on the basis of one's implicit attitudes?[1] Many philosophers say "no," and most parties to the debate agree that the question at least admits a univocal answer. I say: it depends on the type of implicit attitude that produces $X$'s behavior. Even restricting to actions the primary determinants of which are

[1] I mean moral responsibility throughout in the sense of accountability, rather than mere attributability. On the distinction, see Watson (1996), Shoemaker (2011), A. Smith (2012), and Zheng (2016).

implicit associations, sometimes one is responsible for the action in question (blameworthy or praiseworthy, as the case may be), sometimes not. What we need is to explain this difference.

On the one hand, I've argued in Chapters 2 and 3 that implicit biases can seize control of one's behavior—and even practical reasoning and deliberation—from the grip of what one values, cares about, or identifies with to varying degrees. Implicit biases can hijack reasoning itself. And all else equal, when they do, one is less autonomous and less responsible for the resulting actions.[2] In this chapter, I aim to play the other hand: to explain why people *are* responsible for acting on their implicit biases when they are. In particular, I argue that some implicit biases operate via *implicit self-associations (ISAs)*, and that these are "integrated" enough to constitute part of an agent's practical perspective, such that all else equal, the more one's actions express these attitudes, the more one is responsible (and autonomous).

A number of philosophers deny that we can be morally responsible for implicitly-produced behavior precisely because they take implicit attitudes to lack integration (Holroyd 2012, Levy 2014a, 2014b, Smith 2015, and Glasgow 2016). In Section 3, I argue that these philosophers overlook ISAs. In Section 4, I appeal to the philosophical literature on identification that links moral responsibility to integration, and then leverage this connection to argue in Section 5 that one is comparatively more responsible for acting on the basis of one's ISAs than on one's "mere" implicit attitudes. In many other cases, $X$ may not be, but if $X$ shoots $Z$ primarily because of an implicit racial bias that's also an ISA, he is quite blameworthy.

## 4.2 Lack of Integration

Implicit racial biases have been shown to influence a number of real-world discriminatory behaviors, as discussed in Chapter 2. Many of these biases are exacerbated by other morally distasteful associations (see also Chapter 5). Young black males are implicitly perceived as less childlike than white males, for example, increasing the likelihood of the use of violent force against them (Goff et al. 2014). Tamir Rice was 12 years old, holding a toy gun on a playground. In recordings of the incident, responding officers can be heard describing Rice as "maybe 20." 20 year olds, unlike 12 year olds, don't typically play with toy guns. Tamir was tragically shot and killed within two seconds of the officers' arrival.

These are terrible outcomes, but are those who act on implicit biases *responsible* for bringing them about? A number of authors deny as much. Joshua Glasgow (2016) argues that agents are not morally responsible for implicitly-produced actions precisely because they are *alienated* from them (the opposite of identifying). And Neil Levy (2014a, 2014b) and Holly Smith (2015) deny moral responsibility for such actions for much the same reason.[3]

[2]Roughly, I take one to be morally responsible for actions one is autonomous with respect to, but in this chapter I turn attention to the question of responsibility itself. I return to the connection with responsibility in Section 4. I assume that responsibility, like autonomy, comes in degrees—that it's not all or nothing.

[3]More precisely, Glasgow (2016) holds that alienation is exculpatory if one's attitude itself isn't harmful. For an introduction to this literature, see Brownstein and Saul (2016). On responsibility for implicit attitudes generally, see Doris (2002, 2015), Murray (2015), Nahmias (2007), Nelkin (2005), and Vargas (2013a, 2013b).

Holly Smith (2015) argues that we only praise and blame *whole persons*, not just punctate mental states (which is all implicit attitudes can be, according to Smith). It's only the full set of an agent's evaluative attitudes or her *moral personality* that we hold an agent herself morally accountable for.[4] In contrast to explicit judgments, Smith (2015: 200) holds that implicit attitudes tend to be quarantined, relatively isolated from one's other implicit and explicit attitudes, such that the behaviors they lead to typically don't reflect much of the agent's evaluative perspective or anything like her "all-things-considered desires" or judgments. A whole person shouldn't be held blameworthy, for example, for an action springing solely from some rogue gut aversion that happens to reside in her psyche.

Similarly, Levy (2014a: 27) argues that *integration* is necessary for moral responsibility but that implicit attitudes are not (sufficiently) integrated. As Levy notes, both implicit and explicit attitudes are sensitive to semantic content, but only the latter have the propositional structure required for normative, reasons-responsive transitions between mental states. In particular, Levy (2014a: 31) argues that explicit attitudes integrate agents by being employed in reasoning—e.g., believing "If P, then Q" and "P" tends to activate the belief "Q." Implicit attitudes aren't capable of anything like this degree of reasons-responsiveness. They're typically even blind to negation (Chapter 2)—"not poison" still causes implicit gut aversion (Rozin et al. 1990). And two-word phrases cannot be used as primes because each word has an independent, non-compositional implicit effect (Baumeister & Masicampo 2010). Explicit attitudes are sensitive to these and more sophisticated logical and other normative relations between contents, and so are subject to pressures of rational consistency—e.g., pressure not to believe that something both is and isn't poison. Because implicit attitudes are *not* sensitive to (even basic) rational relations (like negation), they're not "integrated" in the sense Levy recognizes, but only "more loosely bound together" (Levy 2014a: 30).[5]

Levy (2014a: 26) goes on to argue that because of this lack of rational integration, implicit attitudes are wholly "encapsulated from other representations." Implicitly-produced actions are "not caused by the system of values constitutive of the agent" (Levy 2014a: 36, echoing Smith on moral personality). As I'll go on to argue in subsequent sections, this is to conflate two senses of integration: there are non-encapsulated values, partly constitutive of the practical agent, other than the rational variety. It's integration in this more inclusive, not-necessarily-reasons-responsive sense that's central to moral responsibility.

Both Levy (2014a: 28) and Smith (2015: 195–6, 206) grant that we might be "indirectly" responsible for implicitly-produced actions—e.g., in the sense that drunk drivers are. Even when one's responsibility-relevant capacities are impaired at the time of action, we blame one for failures to properly exercise those capacities in the *past*, like deciding to get drunk without any other way home.[6] Similarly, one might be indirectly responsible for implicitly-produced actions that one could have prevented in the past. For instance, one might be

[4]Though there are other evaluative judgments we can make about particular attitudes (e.g., 'bad' rather than 'blameworthy'), and we hold whole persons responsible *for* having particular attitudes.

[5]Implicit attitudes are at least not *as* reasons-responsive as explicit attitudes (Levy 2015).

[6]There's of course much more to be said about "tracing" responsibility back to past actions (or omissions). For discussion, see A. Smith (2005), Vargas (2005), and for a response, Fischer and Tognazzini (2009).

indirectly responsible for becoming consciously aware that one harbored some implicit bias but doing nothing to remove it or curb its influence on one's behavior. But, Levy and Smith maintain, we lack *direct* responsibility for most implicitly-produced actions themselves.

Others think indirect responsibility is all that's needed. Jules Holroyd (2012) agrees that integration is not in the offing for implicit attitudes, but argues that indirect, "long-arm control" of the sort involved in the drunk driver case is sufficient for responsibility, which we have over implicit attitudes.[7] However, while some form(s) of long-arm control over implicit biases is indeed possible, much of this is difficult to implement and is typically short-lived (Lai et al. 2016). More importantly, long-arm control simply cannot account for the sort of responsibility agents have for some actions produced by implicit biases.

Stipulate that officer $X$ shoots $Z$ because of an implicit racial bias that $X$ has very little long-arm control over. Suppose that $X$'s action is much more dependent on the bias than on any of $X$'s explicit attitudes, present or past—that only interventions on the implicit level would make any difference to whether the shot gets fired or not. Stipulate that $X$ is not only genuinely explicitly egalitarian, but that he's taken every reasonable precaution to mitigate the influence of his implicit biases (attending supplementary, non-required "de-biasing" training), but despite these supererogatory efforts, all for naught. In the remaining sections, I argue that there are cases of even this variety where $X$ is quite blameworthy.[8]

Foreshadowing, we can borrow a short argument to this effect. Mark Twain's (1885) character Huck Finn decides not to return the runaway slave Jim to Jim's owner because of Huck's implicit attitudes, even though Huck's explicit, all-things-considered judgment is that he should. Nomy Arpaly (2003: 79) argues that Huck and his ilk are *directly* morally responsible for their actions in cases like these, "rather than for any kind of [failures of] self-training or character-building on their parts." If we accept that Huck can be (considerably) directly *praiseworthy* for his action (and I agree with Arpaly we should), there's little reason to think that officer $X$ cannot be just as directly *blameworthy* for his. I return to this argument in closing, where I propose a simpler explanation of the phenomena than Arpaly's.

## 4.3 Implicit Self-Associations (ISAs)

An old folk-psychological chestnut has it that some people racially discriminate because it boosts their perceived self-worth, making them feel racially superior to others.[9] Psychologists have recognized this connection since Allport (1954; cf. Greenwald & Banaji 1995), and recent studies experimentally confirm that threats to one's implicit self-worth make some people more likely to deploy implicit racial biases against others (Kunda & Spencer 2003). This connection is explained by—and constitutes the first strand of several interconnected pieces of evidence for—there being not only "mere" implicit associations, but also *implicit self-associations (ISAs)*: associations between one's concept of oneself (SELF) and other

[7] This grounds responsibility for *acting* on, not *having*, implicit attitudes (Holroyd 2012: 291).

[8] Exerting more long-arm effort may still make $X$ comparatively *less* blameworthy (Arpaly 2003, 2006).

[9] I drop the "perceived" qualifier in what follows; nowhere is the actual worth of individuals at issue.

concepts or "mere" associations (Greenwald et al. 2002, Gawronski, Bodenhausen, and Becker 2007). The association between BLACK and BAD, e.g., is a "mere" association, whereas any further association between it and one's concept SELF is an ISA.

Most people have strong associations between SELF and positive affective valence—they have high implicit "self-esteem" or self-worth (Farnham, Greenwald, and Banaji 1999). When one associates SELF with other concepts or associations, one transfers that affective valence to them.[10] This relationship is also bi-directional, explaining why some employ implicit biases when their self-worth is threatened. Several lines of research bear this picture out.

First, people often "anchor" on the default or status quo and display the endowment and ownership effects (as mentioned in Chapter 3): they require much more money to sell items they already own than they're willing to pay to acquire identical items. Cameron & Ariely (2000), for example, find that people require *14 times* as much to sell NCAA Final Four tickets than they're willing to pay to acquire those very same tickets.[11]

Further evidence suggests these effects are not driven by explicit attitudes or considerations that participants would take to be normative, but instead by ISAs—in particular, work on *post-decision preference shifts*, a variety of (post-hoc) rationalization. Psychologists have known at least since Brehm (1956) that *after* people choose between two alternatives that they initially rate as equally attractive (where the alternatives are usually objects of some kind—like bottles of wine or paintings), they typically rate the chosen as more preferable than the unchosen alternative (cf. Slovic 1975). Many researchers have thought these post-decision preference shifts are the product of cognitive dissonance: the uncomfortable feeling of explicitly, consciously holding two contradictory ideas, and the motivation to revise one by shifting one's explicit beliefs or preferences (Festinger 1957, Festinger & Carlsmith 1959).

Gawronski, Bodenhausen, and Becker (2007) propose an alternative explanation for post-decision preference shifts—the second main strand of evidence for ISAs. They distinguish explicit, System 2 evaluative judgments from implicit, System 1 associative affective reactions that are automatically activated.[12] Cognitive dissonance only occurs at the level of explicit judgments and propositional processes (a felt need to revise *contradictions* cannot apply to mental states incapable of having truth values). Gawronski et al. (2007), however, show that post-decision preference shifts are driven primarily by implicit affective reactions.

Specifically, post-decision preference shifts are subserved by *associative self-anchoring*,

[10]This may be mediated by associations to in-group, family, neighborhood, culture, and other concepts associated with SELF. Sechrist & Stengor (2001) show implicit attitudes are sensitive to peer attitudes.

[11]Kahneman, Knetsch, and Thaler (1991) gave some participants a coffee mug at the beginning of an experiment while others received no mug, and then they asked participants what the minimum was that they'd be willing to sell the mug for or what the maximum was that they'd be willing to pay for it. Participants who already owned a coffee mug required about twice as much to sell the mug as those who hadn't been so endowed were willing to pay to acquire one. Many propose that the endowment effect is driven by people's tendency to value objects they own more than objects they do not, and Greenwald & Banaji (1995) and Morewedge et al. (2009) suggest that the endowment and ownership effects are actually just the same effect.

[12]Departing from Gawronski et al. (2007), associations do not have propositionally structured contents, so they cannot be "evaluative" or involve implicit self-"esteem" in any literal sense. Instead, they exhibit positive or negative valences of varying strengths and can involve one's self-"worth," as I argue below.

the formation of an association between an object and SELF that leads to a transfer of whatever further associations and affective valence one has with SELF to the object.[13] Gawronski et al.'s (2007) results suggest that *choosing* an object can create such an anchor. Their evidence for this is ingenious: Participants with positive implicit self-worth show strong post-decision preference shifts, assigning a higher value to the object chosen (a painting) after the choice.[14] However, participants whose implicit self-worth is negative (i.e., those who are depressed, or associate SELF with a negative valence) display no preference shifts, assigning the same or a lower value to the object chosen as they did pre-choice.

These results show that people implicitly transfer the affective valence (of a particular strength) that they associate with SELF to the objects they choose as a result of choosing those objects (likely to varying degrees, and perhaps weighted according to some function). Zhang & Chan (2009) show that only implicit, not explicit, self-worth predicts the size of post-decision shifts. Other work confirms these preference-shifts are implicitly-driven, as I argued in Chapter 3. Cognitive dissonance would require (accurate) conscious recollection of what choice one made. But preference shifts are observed in young children and capuchin monkeys (Egan, Santos, and Bloom 2007), those with anterograde amnesia (Lieberman et al. 2001), and in preferences for smells after choices have been forgotten (Coppin et al. 2010). Coppin et al. (2012) and Sharot et al. (2012) show that these shifts are long-lasting, and Johansson et al. (2014) show that they affect subsequent choices even after initial "blind choices," where participants are mistaken about which object they in fact chose. As Gawronski et al. (2007) note, this same mechanism explains the endowment and ownership effects: if people develop ISAs with the objects they choose to buy or endow themselves with, and tend to have high implicit self-worth, then associative anchoring explains the tendency to place higher value on what one owns compared to what one doesn't, once one owns it.[15]

Self-associations are implicated in numerous other results, as well, like the *name-letter effect* and *birthday effect* (discussed in Chapter 2): Virginia and Georgia, e.g., are 36% more likely to move to states with those names than others, and people prefer numbers associated with *their* birthdays (Pelham et al. 2002). Participants subliminally induced to associate a number with their name also subsequently make more favorable judgments of women wearing

[13]See Greenwald & Banaji (1995), Cadinu & Rothbart (1996), and Walther & Trasselli (2003). Otten (2003) shows that associative self-anchoring also occurs in the definition of in-groups and out-groups.

[14]In the first instance, choices affect implicit attitudes, which can then percolate up to affect explicit judgments (Gawronski et al. 2007: 229). Note that these one-off choices, then, do seem to constitute one way in which implicit attitudes (perhaps only ISAs) can be formed without repeated, classical conditioning.

[15]Brendl et al. (2005) show that similar effects can be induced by self-affirmation (not just choice), and Perkins and Forehand (2012) show that "pre-existing" ISAs are enough to drive preference shifts and ownership effects in the absence of choice and any primes regarding self-worth. Perkins and Forehand gave participants an IAT asking them to categorize stimuli into classes of self, other, digital clocks, and analog clocks on two response keys. They found that participants showed implicit preferences for whichever type of clock was assigned to the same response key as self-related stimuli. This effect was replicated with fake brand names. Prestwich et al. (2010) show that inducing implicit self-associations with drink A (compared to drink B) increases people's implicit preference for A, as determined by IAT (effects that do not generalize to explicit attitudes, but the size of which varies with level of implicit self-worth; cf. Tietje & Brunel 2005).

shirts with their assigned number compared to those wearing shirts with control numbers (Jones et al. 2002). Indeed—and this is the third main strand of evidence for ISAs—it seems that one of the largest, most consistent patterns in the dual-processing literature is that many of the strongest effects implicit attitudes are capable of involve general tendencies to promote one's own self-worth: the result of *implicit egotism* (see Pelham et al. 2005 and Holland et al. 2009, though this of course doesn't apply without exception).[16] The basic upshot is that ISAs turn out to be major movers and shakers on the implicit level.

The relation between the affective valences associated by an ISA is bi-directional: threats to the worth of what one self-associates with can become threats to the implicit worth of oneself. This explains why threats to one's self-worth can be alleviated by employing implicit biases: if one has an ISA with an implicit racial bias against black people (and so a negative association between one's own worth and the worth of black people), for instance, one can increase one's implicit self-worth by employing the bias and harming or decreasing the *perceived* value of black people. Indeed, Fein and Spencer (1997) find that participants who are told that they have done poorly on a (purported) intelligence test (and so whose self-worth seems threatened) are more likely to attribute stereotypic traits to a gay man than control participants. Moreover, these participants' implicit self-worth is restored when given the opportunity to make judgments expressing these biases, effects also found with implicit racial biases (Sinclair & Kunda 1999, Spencer et al. 1998). A great deal of work now confirms that these mechanisms are behind the folk-psychological chestnut that we began this section with (Fein et al. 2003, Schimel et al. 1999, Sinclair & Kunda 2000, Uhlmann et al. 2010, and Kunda & Spencer 2003). Acting on implicit biases—lowering the perceived worth of those one has biases against—raises one's own implicit self-worth, for some people.

Most importantly, numerous findings demonstrate that ISAs are highly *integrated* in very similar ways to what Levy (2014a) and Smith (2015) require for (direct) responsibility, but roundly deny to implicit attitudes. True, non-SELF-involving, "mere" associations are fairly disunified, and bear no special connection to self-worth. But ISAs all tend to be associated with each other (Greenwald et al. 2002): the name-letter and birthday effects, for instance, are strongly positively correlated with one another (Koole, Dijksterhuis, and van Knippenberg 2001). And implicit associative attitudes also show sufficient internal structure to display the *enemy of my enemy is my friend effect* (Gawronski et al. 2005).

Mandelbaum (forthcoming) argues that the enemy of my enemy is my friend effect cannot be explained associatively: if these attitudes were purely associative, then supposedly disliking $A$ and being aware that $A$ dislikes $C$ would lead you to *dislike* $C$ (even more), since $C$ has been paired with *two* negative stimuli (Walther 2002). Contra Mandelbaum, though, implicit associations come in different strengths and valences (they encode correlations of different degrees and signs). So, one association might encode that, to the extent that $C$ does well, $A$ does worse (they're enemies), and another might encode that, to the extent that $A$ does well, you do worse (you're enemies). But notice: in that case, you've already encoded

[16] Hume (1738/1978) over-stepped in arguing that "self-love" is the strongest source of motivation period, but perhaps he was correct if his claim is restricted to those mental states that do turn out to be associative.

that to the extent that $C$ does well, you do well ($C$ is your friend). In this way, the enemy of my enemy is my friend effect can be explained purely associatively. Contra Mandelbaum (forthcoming), Gawronski et al.'s (2005) findings don't show that implicit attitudes are rationally related; they instead show that a special subclass of implicit attitudes—ISAs—have a more intricate functional role and internal psychological structure than others.

I develop the details of this type of integration in Section 5, but we can already appreciate why ISA's would display it: there are pressures for ISAs to form a coherent perspective on the world on pain of one and the same agent (or self)'s having self-defeating combinations of affective vulnerabilities. Self-associating with both the Yorks and the Lancasters, or any other feuding parties (without some hierarchy), for instance, could lead to fissures in one's implicit self-conception and self-worth. Because ISAs all involve the concept SELF—and there are strong psychic pressures to have a unified SELF, as well as a unified set of affective valences associated with it (pressures that do not apply to other concepts and associations)—we can expect ISAs to form a more unified practical outlook than the rest of one's implicit attitudes—to perhaps constitute an implicit "moral personality" in Smith's (2015) sense.[17]

By way of spelling out this type of integration, I'll begin by arguing for its importance to moral responsibility in the next section. Given what we already know about ISAs, though, and with the help of the parity-of-reasoning argument from Huck Finn, it's easiest to jump ahead to the conclusion now. Compare the following two versions of our main example, holding fixed long-arm control, conscious awareness, and any factors besides ISAs:

> (V1) Despite his genuine explicit egalitarianism, police officer $X$ shoots $Z$ because of an implicit racial bias that he also has a strong positive ISA with. Assuming the bias is sufficiently associatively integrated into $X$'s other ISAs and he associates SELF with positive affect, he'll tend to feel self-satisfied and racially superior as a result of acting on his implicit racial bias—doing so will tend to boost his implicit self-worth, such that $X$ implicitly feels good about himself for killing $Z$.
>
> (V2) $X$ shoots $Z$ because of an implicit racial bias, but not one that $X$ has any ISA with. In this case, the implicit bias that $X$ acts on bears no connection to his implicit self-worth and he feels no satisfaction or superiority in acting on it.

In (V1)—especially (V1), and not in many other possible variations on the case—I believe $X$ is quite (directly) responsible, blameworthy for killing $Z$. Again, if we reject this conclusion, we must also reject that Huck Finn is responsible in the cases of interest to Arpaly (2003).

In addition, the comparison provides some initial purchase on what responsibility depends *on* in these cases: Officer $X$ is at least much *more* responsible in (V1) than (V2), but the only difference between these variations is the presence or absence of an ISA with the implicit

[17] Again, though, this need not be *evaluative* or involve *esteem* in any sense necessitating propositionally-structured content, and so perhaps would be better termed an implicit "affective personality."

bias.[18] Hence, ISAs have a large intuitive impact on moral responsibility. Returning to our initial question, we reach the following answer: if $X$ shoots $Z$ because of an implicit racial bias that also constitutes an ISA, or that $X$ has an ISA with,[19] $X$ is more blameworthy for doing so compared to when he acts on the same implicit bias but has no ISA with it.

I agree with Levy and Smith that one is more responsible for actions the more integrated their mental causes. And I agree that implicit attitudes—across the board—tend to be considerably rationally unintegrated. However, the evidence in this section suggests that it's wrong to deny *all* implicit attitudes *all* types of integration. ISAs are unique in exhibiting an associative, *arational* variety of integration. To elaborate on the details of this type, we can turn to accounts of why integration is important to moral responsibility generally.

## 4.4 Integration and Moral Responsibility

Levy (2014a: 36) and Smith (2015: n. 44) are explicit that their appeals to integration—and the attendant insight that responsibility, praise, and blame are assessments of whole persons—are motivated by theories of identification. Harry Frankfurt (1971) introduced the notion of identification to overcome shortcomings of "classical compatibilism," according to which (rone is responsible for an action just in case it's produced by one's internal mental states rather than external factors. Incompatibilists (libertarians, in particular) claim this fails to account for the fact that we're not responsible for some internally-produced actions. Agents can also be constrained or coerced by their own internal mental states, and incompatiblists claim these phenomena require a homoncular *causa sui* to endorse (or disavow) one's mental states. Frankfurt introduced identification as the extra structure compatibilists need to distinguish internally-produced behaviors we're responsible for from internally produced behaviors that we're not, without recourse to such metaphysical extravagance.

*Identification* is a matter of "where (if anywhere) the person himself stands," where this standpoint is a matter of having further attitudes of a specific kind, not something mysteriously hovering "over and above" the entirety of one's mental states (Frankfurt 1987: 166). As Christine Korsgaard (1996: Ch. 3) puts it, when we have contradictory inclinations or initial desires, we're forced to "step back" from and adjudicate between these. The mental states we form as resolutions on how to proceed are constitutive of identification.

Frankfurt (1971) distinguishes first-order desires from second-order desires for first-order desires to lead one to action (*volitions*). He imagines a willing and unwilling addict: both have a first-order desire to take a drug, but while the willing addict has a volition for this desire to be effective, and so identifies with it, the unwilling addict has a volition not to take

[18] $X$ is even less responsible than in (V2) if, rather than lacking any ISA altogether, he has the "opposite" ISA: a *negative* association between SELF and the implicit racial bias—a "dissociation" or implicit self-alienation from the bias, such that he feels self-loss or diminishment as a result of acting on it.

[19] If ISAs are associations between SELF and other singular concepts, like BLACK, then implicit biases themselves might constitute ISAs. If ISAs are "higher-order" associations—between SELF and "first-order" or "mere" associations, like BLACK/BAD—then one can only have an ISA *with* such an implicit bias. I can't see that the empirical results yet adjudicate which structure(s) ISAs possess, if not both.

the drug (he's "alienated" from his desire for it). This is the distinction compatibilists need: even though both addicts act on the basis of an internal desire to take the drug, the willing addict is more responsible than the unwilling because only he identifies with his desire. In this way, identification theorists hope to explain freedom and moral responsibility in terms of autonomy, or self-guidance, where guidance and governance of actions *by the agent* is understood in terms of guidance of the agent's actions *by particular attitudes* of the agent.[20]

For Korsgaard (1996), self-guidance mainly amounts to answering the question "what to do?"—resolving the impasse between initial inclinations once one steps back. These answers are both determined by, and determinative of, one's *practical identity*: "the description under which you value yourself, a description under which you find your life to be worth living and your actions to be worth undertaking" (Korsgaard 1996: 101). Practical identities include one's deepest values, roles, and practical commitments—e.g., with respect to one's career, spouse, "ground projects" (Williams 1981), and, for Korsgaard, commitment to the moral law. For Korsgaard, one identifies with what's included in one's practical identity.

Michael Bratman (2007: 40–2, 95–102) worries that the self-reflexive content of practical identities alone won't secure their inclusion in "where one stands." Thinking *about* oneself as being some way doesn't necessarily make one that way. What's in fact central—which Bratman is perhaps most clear about—is the functional role of the attitudes in question.

Bratman (2007: 52) takes identification to consist of *self-governing policies* (intentions) to treat given desires as reason-providing in behaviorally efficacious practical deliberation (and being satisfied with these policies—not having others that interfere with them).[21] Unlike Frankfurt, Bratman isn't tempted to eschew normative content entirely: higher-order desires for lower-order desires to lead one to action are not enough. Desiring that an irresistible first-order urge move one to action by bypassing one's practical reasoning entirely isn't identification. Bratman (2007: 38) claims that one must also take one's first-order desires to provide justifications or reasons of various weights in thinking about what to do. The self-governing policies or *valuings* that assign various weights to one's desires in practical deliberation have subjective normative authority to "speak for the agent" because they help to constitute the very perspective from which the question of what's valuable arises.[22]

The unique functional role played by the attitudes constitutive of identification lies in the further organization they confer on one's mental economy—e.g., by differentially weighting each of one's first-order desires in deliberation. This type of organization is required for having a perspective or standpoint unified enough to qualify as that of one and the same persisting agent in the first place—the sort of causal unit it makes sense to praise or blame.[23]

[20]Frankfurt also introduces *wantons*, who have no second-order desires whatsoever and do not care which first-order desires "win out." Says Frankfurt, wantons are not even persons. Frankfurt (1999: 104–5) goes on to require that a volition be *wholehearted* to count toward identification—be had in "the absence of any tendency or inclination to alter its condition," such that wholeheartedness isn't a matter of having any particular attitude(s), but a "state of the entire psychic system." More on wholeheartedness in Section 5.

[21]Compare satisfaction to wholeheartedness. Policies must also be reflexive and non-instrumental.

[22]Bratman (2007: 25, 105); compare this weighting in deliberation to Korsgaard's (1996) "stepping back."

[23]Bratman maintains the neo-Lockean thesis that identification partly constitutes personal identity over

It's widely agreed that identification plays a crucial role in the unity of the practical agent because it's a central way of binding one's own well-being up with that of other things or beings. Frankfurt (1982, 2002a: 187-8, 2006a: 61), for instance, now stresses that love and what one *cares* about are the constituents of identification. Failure to satisfy a mere first-order desire isn't a *loss* in the way that failure with respect to what one cares about is:

> A person who cares about something is, as it were, invested in it. He *identifies* himself with what he cares about in the sense that he makes himself vulnerable to losses and susceptible to benefits depending upon whether what he cares about is diminished or enhanced" (Frankfurt 1982: 83).[24]

Agnieszka Jaworska (2007) and Chandra Sripada (2016, forthcoming) also offer accounts of identification centered on caring. Attitudes such as these constitute where an agent "stands" because they form a perspective that helps determine how well (or poorly) things go for the (whole) agent. They're constitutive of what the agent takes to be personal losses and gains, and so are under pressure to be integrated with each other. "In extreme cases," acting on a desire that goes against one's identification is to act against one's will, in the sense that "this would amount to a kind of self-betrayal or a *failure* of self-respect or *self-esteem*" (Bratman 2007: 161, my emphasis).[25] That these attitudes determine what count as losses and gains for an agent (how well or poorly her life goes) explains why moral responsibility properly accrues to actions that are (non-deviantly) produced by, and so express, these attitudes.

The parallels with the data discussed in the last section should by now need no reveal. On first pass: just as there are first-order and second-order desires, so there are "first-order" (non-SELF-involving, or "mere") implicit associations but there also "second" or "higher-order" implicit self-associations (ISAs). And just as agents are more morally responsible for acting on second-order compared to first-order desires, so agents are more responsible for acting on ISAs compared to mere "first-order" associations.[26] What's important is that ISAs play the same basic functional role with respect to non-SELF-involving associations that Frankfurt's volitions are meant to play with respect to first-order desires. It's this (type of) functional role that's relevant to autonomy and responsibility, or so I'll argue.

ISAs are near-ringers for Korsgaard's (1996) practical identities: they are *associations* (rather than propositional descriptions) "under which you value yourself," "under which you find your life to be worth living and your actions to be worth undertaking." ISAs are not implicit judgments and tend not to respond to reasons, but they do constitute an affectively unified perspective on the world—a coherent way of caring about or being emotionally vulnerable to and bound up with various bits of it. As such, ISAs partly determine how

time. Velleman (2002) criticizes such proposals, but even if identification isn't necessary for strict numerical identity, there is something like "who one is," practically, or one's "moral personality" that it is central to.

[24] "What one cares about is measured by how much one is 'invested' in or bound up with something, by one's sense of loss or diminishment upon not realizing or achieving the object" (Watson 2002: 148).

[25] Frankfurt (2002b: 277-8, 2006b: 16–18) is also explicit that self-esteem lies at the core of identification.

[26] Calling associations first- and second-order may be misleading (as it may be with desires). See n. 19.

well one's life goes. As with explicit identification (Frankfurt 2002a, 2006a, Bratman 2007), diminishment of what one has ISAs with involves felt loss and diminishment of oneself.

ISAs, though, are characterized precisely in terms of their functional role, addressing Bratman's challenge to Korsgaardian accounts. We understand much about the mechanisms involved in their formation (associative self-anchoring), their internal structure (involving SELF), and their downstream effects (in the ownership, endowment, name-letter, implicit egotism, and many other well-replicated effects). We also know much about their elicitation, and in a way that's tightly coupled to responsibility: threats to implicit self-worth and the consequent drive to restore it often lead people to act on their ISAs. In other words, ISAs are intimately bound up with (one's own sense of) how well one's life is going. Empirical evidence also confirms that ISAs are more integrated than other implicit associations.

However, different theories of identification differ over just what type of integration is relevant to moral responsibility. And while the non-explicit nature of ISAs makes them foreign to all traditional theories of identification, which are squarely restricted to explicit attitudes, ISAs are more easily accommodated by some theories than others. The type of integration displayed by ISAs may not be exactly identical to the type exhibited by any variety of explicit identification, but in the next section I argue that ISAs do play the same basic functional role within one's mental economy as some attitudes traditionally implicated in identification (especially caring)—just on the implicit, rather than explicit, level.[27]

## 4.5 Associative Integration

Many if not most identification theorists take conscious, reflective endorsement to be essential to identification. There's reason not to. Unlike the simple possession of certain attitudes or the obtaining of certain structures or relations between them, endorsement invites back in worries about a *causa sui* floating outside the entirety of one's mental economy—something over and above them that has to *do* the endorsing of any particular attitude.[28]

In a related vein, following Gary Watson (1975), many take identification to consist in evaluative judgments—explicit beliefs about values or reasons, not just "valuings" (which can have desire-like direction of fit, instead). Like all beliefs, one's evaluative beliefs must

[27]The claim is not that any set of ISAs taken in isolation would constitute a practical agent. Explicit forms of identification may be necessary for possessing the *capacities* for agency and moral responsibility. The claim is instead that *given* an agent in possession of (not necessarily exercising) whatever capacities these are, ISAs are partly determinative of whether and how morally responsible that agent is for any particular action because the products of this implicit "affective personality" accrue to the possessor of those capacities.

[28]It's for this reason that I believe we should be wary of construing identification in terms of a "real self," as Wolf (1990), Sripada (2016), and others are inclined to. Reflective endorsement may also invite onto the scene a familiar vicious regress: how can one's endorsement constitute "where one stands" if one didn't endorse *it* (the first endorsement, at a higher-order)? (Arpaly & Schroeder 2012; cf. Watson 1975). Some identification theorists are willing to give up the conscious reflection requirement, like Frankfurt (2006b: 6), who claims that "[h]igher-order responses need not be especially thoughtful, or even entirely overt." Bratman (2014: 105; cf. 2007: 6, 192) also accepts that "certain less demanding social psychological phenomena might in certain cases to some extent functionally substitute" for explicit self-governing policies.

be logically consistent with one another, so there's pressure on them to constitute a more unified standpoint than the heterogenous motley of one's mere desires. Levy's (2014a: 31) account of integration in terms of inferential and logical relations is such an account. This variety of integration—rational consistency—may well be necessary to have the *capacity* for moral responsibility. But it's certainly not the type of integration displayed by ISAs.

Contra Holroyd (2012) and Uhlmann, Brescoll, and Machery (2010), people do not act on their ISAs *in order* to boost their implicit self-worth—i.e., because they take there to be reason for performing actions that have that effect. Associations—ISAs included—lack propositional structure, and so cannot involve taking anything to be a reason at all, even implicitly. One does not act on them in order to satisfy an end. Associations cannot represent that something "is good" or "is wrong" or *is* anything. Of course, ISAs and other implicit attitudes can have both upstream rational causes, as when someone else's testimony produces certain associations, and they can have downstream effects on reasoning. But implicit attitudes themselves are only directly involved in associative mental transitions.

Instead, associations directly attract one toward, or repel one away from, different actions. Actions motivated by ISAs *express* one's implicit valuings (to a greater or lesser extent) in a way that's intelligible affectively and associatively, not rationally. There is, for instance, a certain arational sense or "logic" to the way some emotions wash into others—e.g., how fear of predators can become fear of the dark, or positive feelings toward one person can bleed into positive feelings for that person's friends. Hursthouse (1991) argues there are "arational actions" that merely express emotions or desires, without being done for the sake of those desires or taking them to provide or represent any reasons or justifying considerations. We might be able to explain why a man is rolling around in his recently-deceased spouse's clothes in terms of what first-personal sense it makes to him to do so—referring to his grief, for instance. But this doesn't mean he takes his grief to provide a reason for so acting, or that he acts in order to grieve.[29] ISAs involve coherence in this type of intelligibility, not the intelligibility of inferential consequences or normative consistency. So the type of integration displayed by ISAs is not a type that theories of identification requiring reflective endorsement or evaluative judgments can countenance as relevant to moral responsibility.

However, the more traditional take on integration has never required specifically rational coherence, as Frankfurt has always made clear. Rational consistency constraints on beliefs are one form of integration, but there are others—e.g., pressures toward unity that come from the consistency constraints on *actions*: that actions not be *self-defeating* (Frankfurt 1999: 99). Attitudes with desire-like direction of fit and whose downstream influence extends all the way to action—like intentions and volitions, but unlike first-order desires—inherit these consistency constraints on action, given their functional role in guiding it (Bratman 2007: 68-88; compare Frankfurt 1999 on wholeheartedness). Having first-order desires to vacation in Hawaii and Alaska (at a given time), even though one cannot do both, is not practically

[29]There are explanations of why some actions happen that can be given in terms of what sense there was in doing them from the agent's own point of view that are not explanations of what reasons the agent took there to be for performing them. One might automatically shout at a stranger only because the stranger resembles one's nemesis (an associative relation), even though one sees no reason at all to shout at anyone.

self-defeating or inconsistent. One and the same practical agent's *intending* to do both is.[30] Conflicting ISAs are also incoherent in this self-defeating sense, just not rationally so.

ISAs are no more intentions or volitions than they are explicit judgments, and they're not under the same consistency constraints as either. But ISAs are subject to the same more general pressure not to pull one and the same agent asunder. "Seeing" oneself as both "someone who always zigs" and as "someone who always zags" or as equally a Hatfield *and* a McCoy, is self-defeating, even where these "seeings" are only associative attractions and repulsions. The empirical evidence shows that ISAs are sensitive to this broad kind of (in)coherence—e.g., to not being enemies with the enemy of one's enemy. The coherency constraints on ISAs follow the logic of affective self-attachment, rather than theoretical or practical rational consistency, but they arise from the same broad functional role as the coherency constraints on intentions and caring (though direction of fit is somewhat blurred on the implicit level). ISAs are under pressure not to involve self-defeating combinations of affective vulnerabilities and sets of attractions toward and repulsions from actions.[31]

What's essential to identification is sufficient integration into a practical perspective unified enough to qualify as the standpoint of a single practical agent. But there are different types, or realizers, of integration (we should be "pluralists" about it).[32] Given beliefs' function of representing the world, they're under rational pressure not to be theoretically inconsistent with each other, on pain of failing to fulfill this function. Given intentions' function of leading an agent all the way to action, they're under rational pressure not to be practically inconsistent, which requires non-self-defeating intentions and a hierarchy of ends. Given the function of *ISAs* (like what one explicitly cares about) in determining one and the same agent's affective attachments to other things, beings, and actions, ISAs are under pressure to be associatively co-intelligible, not pulling SELF in opposite directions (which may also require a hierarchy of ISAs).[33] All these types of attitude play a role in the unity of

[30] Though see Kolodny (2005) for development of an error theory for many rational consistency constraints.

[31] As a contingent matter, associative self-defeat also tends to diminish positive implicit self-worth, creating additional pressures for ISAs to be integrated. But what's central is that SELF be associated with a unified set of concepts and affective markers, whether positive or not. Not everyone has positive implicit self-worth. In those cases, ISAs still organize the rest of one's implicit mental economy, but they integrate it around whatever diminished or oppositely valenced affect one associates SELF with. Such sets of ISAs may still provide coherent affective personalities or outlooks on the world, even if these are pathological in some sense.

[32] On related versions of pluralism concerning agency, see Vargas (2013a) and Doris (2015: 171–7).

[33] There are three potential dimensions of integration: *temporal* integration or stability over time, *vertical* integration of the sort involved in prioritizing ends and "nesting" plans hierarchically, and *horizontal* integration of the kind exemplified by synchronic rational consistency constraints. I've focused on horizontal integration, but ISAs also exhibit significant vertical integration. Evidence suggests that, like self-governing policies, ISAs may play a role in *weighing* the influence of different "first-order" associations in determining action. Glaser & Kihlstrom (2006), Park, Glaser, and Knowles (2008), and Devine (2002) find that implicit attitudes can inhibit competing implicit attitudes and respond to variations in the means needed to realize their "ends," mechanisms in which ISAs may be implicated. ISAs with more particular things also tend to be mediated by more "fundamental" ISAs—e.g., one self-associates with things because they're associated with one's in-group and self-associates with particular colors or sports teams because of the city or neighborhood one lives in or the school one attends, rather than the other way around. Thus, there may be a hierarchy of

the practical agent through coherency-involving functional roles; their differences arise only from their differing directions of fit and the types of mental transition they're involved in.

Of course, there are those who will insist on rational integration *alone* for moral responsibility. But the evidence suggests that associative integration is similar enough to the type at the heart of many identification theories to make agents comparatively more responsible for acting on ISAs than on other implicit attitudes. We might imagine what it's like for officer $X$ to perform his action when he has an ISA (V1) from the first-person. He sees $Z$. He raises his gun. The prospective imagery of standing over $Z$'s body causes a warm, swelling, proud feeling in him, a sense of the superiority of "$X$'s people" over "$Z$'s people." He feels an implicit attraction to the image of squeezing the trigger and having $Z$ die. He fires... But of course, even for $X$, these vague first-personal impressions are just whatever percolates up from what we've seen, in this section and the last, to be going on under the hood, under the radar of most conscious awareness: $X$'s action expresses what's become a deep-seated feature of who he implicitly is—part of what constitutes personal gains and losses for $X$.

Space precludes rehearsing the full debate with those who insist on rational integration for moral responsibility.[34] By way of closing, however, we can assess how theories based on self-defeat fare against those based on reasons-responsiveness on the implicit level, specifically.

## 4.6 Conclusion

All else equal, the more officer $X$ is motivated by a racial bias that he implicitly self-associates with and the more associatively integrated with the rest of his ISAs that bias, the more he's morally responsible for shooting $Z$, because the more his action expresses $X$'s affective personality or implicit practical identity—who $X$ implicitly is and where he stands on fundamental practical matters.[35] More generally, a person is more directly responsible for an action the more it expresses what he or she self-associates with. One is praiseworthy or blameworthy as these associations are morally appropriate or inappropriate.

What happens when explicit, reflective identification and ISAs conflict—when who one is on the System 1 and System 2 levels come apart? Perhaps this always involves less autonomy and responsibility than "full harmony" between the implicit and explicit levels, and perhaps some further "two-way process of integration" would render one *most* responsible (Friedman 1986). But this may be the exception. The multiplicity of ways implicit and explicit thoughts can conflict limits the refinement of any answer, and further (empirical and conceptual) work is needed. But we can address two of the larger issues in this vicinity in closing.

---

ISAs on the implicit level similar to those posited by Bratman and Frankfurt at the level of explicit thought.

[34]On the basic debate between "mesh theories" and "reasons-responsive theories," see McKenna (2011). See also Watson and Frankfurt's exchange in Buss & Overton (2002), Wallace (2014), and Sripada (2016). Note again that this is not to deny that reasons-responsiveness and rational consistency are necessary for the capacity for moral responsibility, nor that the latter constitutes one type of identification (see note 27).

[35]On self-expressive theories of moral responsibility, see Ch. 1 and Sripada (2016, forthcoming).

First, I argued in Chapter 2 that hijacking of one's deliberation and actions away from one's (explicit) values by implicit attitudes mitigates autonomy (and so also moral responsibility). ISAs require a qualification to this thesis, since ISAs themselves constitute (implicit) values. Assuming that implicit attitudes of both kinds can hijack deliberation and actions away from what one explicitly values and cares about, all else being equal, one's moral responsibility is at least *less* mitigated (one is less "off the hook") when ISAs hijack one's reasoning compared to when "mere" non-SELF-involving implicit attitudes hijack one's reasoning, since the resulting actions in the former but not the latter case still express more of what one values and cares about. This is not to deny that hijacking by ISAs still mitigates moral responsibility to some degree (compared to cases in which no hijacking occurs).

Second, is *X* ever *more* blameworthy than his *explicitly racist* police officer counterparts—those who perform the same actions but are motivated by explicit racism? Nomy Arpaly, sometimes with Timothy Schroeder, addresses the other side of this question in her discussions of Huckleberry Finn and the runaway slave Jim, who travel together up the Mississippi (see Section 2). As their adventures progress, Huck and Jim bond, becoming close friends. When Jim is captured by a family who plans to sell him back into slavery, Huck's explicit attitudes and what he explicitly identifies with unambiguously tell him that the morally correct thing to do—what *God* wants him to do—is to ensure that Jim is returned to his owner, Miss Watson. But when it comes time to decide, Huck finds that he simply cannot do it, and he resolves once and for all to "take up wickedness" and to help Jim escape.[36]

Arpaly & Schroeder (1999) call the type of phenomenon displayed by Huck *inverse akrasia*: Instead of acting against one's explicit judgment and having this lead to an immoral or irrational action, Huck acts in an (objectively) morally better way than he would have had he *not* been akratic, since *his* reflective, all-things-considered judgment is the morally wrong judgment. Arpaly (2003: 78) admits Huck is imperfect, "but as he is, he is better than many, including his counterpart who is liberal in conviction but not in deed."

I agree with Arpaly that Huck is praiseworthy, and more so than some of his counterparts whose heads, rather than hearts, are in the right place. However, the account developed in

[36] So I got a piece of paper and a pencil, all glad and excited, and set down and wrote:

Miss Watson your runaway n—— Jim is down here two mile below Pikesville and Mr. Phelps has got him and he will give him up for the reward if you send. HUCK FINN.

I felt good and all washed clean of sin for the first time I had ever felt so in my life, and I knowed I could pray now. But I didn't do it straight off, but laid the paper down and set there thinking... And got to thinking over our trip down the river; and I see Jim before me, all the time; in the day, and in the nighttime, sometimes moonlight, sometimes storms, and we a floating along, talking, and singing, and laughing. But somehow I couldn't seem to strike no places to harden me against him, but only the other kind... and at last I struck the time I saved him by telling the men we had smallpox aboard, and he was so grateful, and said I was the best friend old Jim ever had in the world, and the *only* one he's got now; and then I happened to look around, and see that paper.

It was a close place. I took it up, and held it in my hand. I was a trembling, because I'd got to decide, forever, betwixt two things, and I knowed it. I studied a minute, sort of holding my breath, and then says to myself:

"All right, then, I'll *go* to hell"—and tore it up (Twain 1885: Ch. 31).

preceding sections of this chapter already provides a very different explanation of these facts. Arpaly (2003: 132, 138–9) takes Huck to show that identification and integration are not central to moral responsibility—*because*, like Levy and Smith, she assumes that identification and integration must be explicit, reflective, conscious affairs.[37] In contrast, I think that Huck is only morally praiseworthy if he acts on the basis of an ISA, for several reasons.

First, Arpaly's is a reasons-responsive account of responsibility. Despite Huck's faulty explicit moral beliefs, she claims that in saving Jim he responds to the objective (*de re*) moral reasons, albeit implicitly—specifically, to Jim's moral personhood.[38] Indeed, Huck need not have any mental states with (*de dicto*) moral content (employing evaluative or normative concepts) at all. Moral responsibility depends solely on the degree to which one is causally sensitive to the objective, *de re* moral reasons (*qua* reasons), consciously or otherwise.

To my mind, the notion of reasons-responsiveness becomes slippery once watered down this much—to potentially "ballistic" brute causal sensitivity (Arpaly 2006: 19). This seems like "reasons-sexing" (after "chicken-sexing" in epistemological debates over reliabilism). All that matters is that one reliably get it right, not at all *how* one sorts the morally good from the morally bad, or whether one has any idea of how one does so. It's dubious that reasons-responsiveness in this sense is a reasonable basis for ascriptions of moral praise and blame.

Worse yet, like all reasons-responsive theories of moral responsibility, Arpaly's account seems to run directly afoul of the abundant evidence showing that most implicit attitudes are highly *un*responsive to (objective) reasons. That extreme unresponsiveness to (moral) reasons is, after all, part of why research on implicit bias has been so captivating, and so often disturbing. Implicit biases are shocking not only because we were, until recently, relatively unaware of them, but precisely because they respond to what most people nowadays would never mistake for reasons or justifying considerations at all, on any level. Most people now would not take race to be a justifying consideration for any kind of violent force, for instance. Arpaly's account simply is not easily applied to most implicitly-produced behaviors.[39]

Arpaly (2003: 84–93; esp. 87–88) does take one's *degree* of responsibility to depend on the "depth" of a motivation's reasons-responsiveness. To be *very* praiseworthy, Huck's responsiveness must be relatively deep, which Arpaly equates with the strength of his motivating attitude in the face of conflicting motivations—how often it'd win out over other motivations counterfactually. However, this creates a dilemma for Arpaly's account: either her theory builds integration back in to the notion of "depth" (in which case it's not an alternative to accounts based on integration), or it suffers serious counterexamples.

Suppose depth requires no particular internal organization, that it's just brute counterfactual strength. Then in principle a mere collection of isolated modules—a "Blockhead"

[37]In fairness, Arpaly's sights are most focused on reflective endorsement, which I agree shouldn't be a requirement on *all* types of identification for partly independent, regress-related reasons (see note 28).

[38]Arpaly remains officially neutral on normative ethics, but tends toward examples with Kantian flavor. Huck's "reluctance [to turn Jim in] is to a large extent the result of the fact that he has come to see Jim as a person, even if his conscious mind has not yet come to reflective awareness of this" (Arpaly 2003: 77).

[39]In other words, I agree with Arpaly about the potential for implicit moral responsibility, but with Levy (2014a, 2014b) and Smith (2015) about the reasons-unresponsiveness of implicit attitudes (see Section 2).

(Block 1981), or a thermostat, could be (deeply) reasons-responsive. Perhaps responding to some reasons would require a psyche with more internal complexity, but if this builds integration back into the picture, it's no longer a real alternative (and integration would be doing all the work—accounting for most of the variation in degrees of responsibility).[40]

Some manipulation cases corroborate the insufficiency of brute counterfactual strength. Consider a case posed by Arpaly herself, in which you're kidnapped by an evil neurosurgeon:

> Imagine that you wake one morning with a desire to murder the Canadian Minister of Sports and Recreation, and nothing else has changed in your brain. If you are anything like me, you will find yourself with a desire radically lacking in *integration* (Arpaly 2003: 167, my emphasis).[41]

That is, a desire might always win out in producing behavior even if it's completely isolated from the rest of one's mental states within its own module. Someone who's otherwise racist, for example, might be manipulated into being a "failsafe egalitarian"—whenever in relevant situations, a quarantined desire might hijack the rest of his reasoning from the control of his biases to produce egalitarian behavior. Such a desire would be "deep"—if depth doesn't require integration—but intuitively, the failsafe egalitarian isn't praiseworthy; it's just a desire *in* him, not to *his* credit that he acts as morality requires. So Arpaly (2003: 132–9) may need to lapse back into more agreement with identification theorists than she'd like.

Finally, in positing questionable reasons-responsive mechanisms to explain phenomena already accounted for using simpler, independently verified psychological states—ISAs—Arpaly's account appears to over-intellectualize. Indeed, Arpaly (2003: 77) herself initially puts the point more simply: Huck has become Jim's *friend*. But we can accept that explanation without jumping to implicit reasons-responsiveness, let alone to moral personhood.[42]

Love and friendship and the caring constitutive of them often grow out of ISAs. Plausibly, Huck has come to implicitly care about Jim by the time he decides to help him escape—he's developed an ISA with Jim and affectively entangled their well-being.[43] The thought of things going poorly for Jim, then—and, in particular, of Huck's being complicit in their so going—makes Huck feel prospectively bad about himself and makes his life seem less worth living (by Huck's own lights), associatively repelling him from helping to return Jim to slavery. I contend that it's this repulsion from diminishing Jim's and his own intertwined implicit worth or well-being that Huck is praiseworthy for. Indeed, assuming his behavior is *implicitly* produced, I think the foregoing worries about Arpaly's account suggest that *unless*

[40] See also Harman (2007). Arpaly & Schroeder (2014: 63, 204) may sneak integration back in by two routes—they note that: (i) rationalizing explanations of action may need to invoke large sets of beliefs and desires and (ii) that if there are multiple fundamental moral goods or reasons, then praiseworthiness requires not only that one's desires be of particular strengths, but also that they have the right *comparative* strengths.

[41] "That is, your desire will be so essentially opposed to the rest of your attitudes, including your deepest ideas and concerns, that it would make the offending desires of the most unwilling heroin addict or kleptomaniac look positively wholehearted" (Arpaly 2003: 167).

[42] Indeed, it's not clear Huck's attitude toward the personhood of black people in general changes at all.

[43] SELF and JIM are strongly positively associated: if things go well for one, they go well for the other.

Huck's action expresses a positive ISA between his own well-being and Jim's, he is not (very) responsible for it—even if his motivation for action is reasons-responsive in some "ballistic," quarantined sense. In that case, far from being an objection to theories of identification based on integration, Huck turns out to be the poster-boy for a new species of it.

On *this* reconstruction, Huck is morally "better than many, including his counterpart who is liberal in conviction but not in deed." And this applies equally to Huck's "inverted" analogues who have "wonderful convictions" but are "immoral in their actions and emotions" (Arpaly 2003: 78). Officer $X$ is just such an analogue: he has explicit thoughts that are perfectly egalitarian, but implicit racial biases. By parity of reasoning, then, implicit racists like $X$ are sometimes surprisingly *more* responsible for their actions than their explicitly racist counterparts.[44] When Huck or officer $X$ acts on a non-SELF-involving implicit attitude, neither is typically very responsible. Holding everything else fixed, when they have ISAs, both are considerably more responsible, potentially more so than their explicit counterparts.

For parity, we must hold everything else fixed and imagine that $X$'s implicit self-worth is just as intimately tied to his implicit bias as Huck's is to Jim's well-being. That's very tightly indeed. So described, though, $X$ will tend to feel deeply (implicitly) superior and better about himself for diminishing the perceived worth of black people—e.g., by shooting $Z$—and he'll tend to feel self-loss and diminishment if he fails to act on this bias (or acts on contrary, egalitarian attitudes). Again, in *this* version of the case, and not in many others, I believe officer $X$ is quite morally responsible—blameworthy—for killing $Z$, because his action expresses to a high degree where he implicitly stands on important practical matters. The attitudes that motivate his action are central constituents of who he, *qua* practical agent, implicitly is. They're part of what makes his life go well or poorly, by his own lights.[45]

Of course, it would be misleading to say officer $X$ is racially biased or that Huck identifies with Jim and leave it at that. We must disambiguate: Huck identifies with Jim on the implicit but not explicit level. He is not "wholehearted" in at least one sense. However, this fissure *between* the System 1 and System 2 levels doesn't entail any *on* either: Huck is unified within each—both a racist *and* a good boy. It may be that "[w]hen we think of ourselves, we identify with System 2" (Kahneman 2011: 21). But this is partly wishful thinking. We are also, in part, our implicit self-associations. In some cases—especially for officer $X$—this is a bitter truth to swallow. But in others—like Huck's—the fact that we, as practical agents, have both implicit and explicit halves, shows us at our best, or at least most human.[46]

It's by now clear that implicit attitudes can have momentous effects on real-world behavior. But so far, we've focused on laboratory-based studies that tend to look at implicit attitudes in relative isolation. The next two chapters move on to look at how these effects play out in actual ecological contexts outside the lab, and how implicit attitudes can interact with one another. Implicit biases are implicated in highly salient problems like police violence, but also more deeply-rooted social woes such as the persistence of chronic poverty.

[44]Parity between praise and blame is a further claim, which some deny (e.g., Wolf 1990 and Knobe 2003).
[45]The difficulty of determining the facts in any individual case should always give pause to actual blam*ing*.
[46]Doris (2015: 160–2) takes schisms of the sort faced by Huck to be typical of the human condition.

# Chapter 5

# Implicit Bias in Cycles of Poverty

## 5.1 Introduction

Implicit biases and other attitudes have captivated academic and public attention in recent years, especially their role in police discrimination. But what do implicit attitudes look like situated in broader ecological context? How are they involved in interpersonal interactions in ordinary, day to day life? To hone in on one particularly important context: what is the role of implicit attitudes in perpetuating poverty—in particular, the cycles of poverty endemic to urban life in the United States? This chapter aims to address that question.

The notion of a vicious *cycle of poverty* is an old one, especially in the context of the American ghetto.[1] St. Clair Drake and Horace Cayton (1945/1993: 175, 268) argue that (explicit) white racial discrimination is responsible for segregating black Americans into high-poverty neighborhoods—conditions which then reinforce and exacerbate white Americans' perception of blacks' inferiority, and in turn increase their discrimination-based segregative practices (see also Myrdal 1944). Through this mechanism of "circular reinforcement," white discrimination confines urban blacks to ghetto neighborhoods. Because of the relative poverty of blacks, ghettos then become very visible signs of perceived inferiority, the substandard health, sanitary, safety, and aesthetic conditions poverty leads to becoming associated with their residents. Blacks can usually only move into areas that are already dilapidated, and tend to be blamed for the conditions of the area, such that these conditions become further rationalization for quarantining blacks behind the "color line." Kenneth Clark (1965: 156) extended this idea to a vicious cycle in multiple arenas of life, with each one reinforcing the others in a "tangle of pathology."[2] Clark's work especially highlights that there may be no *one* cycle of poverty, and in the final analysis, all parts need to be understood together.

[1] On the similarities and differences between (European) Jewish ghettos at different historical periods and (overwhelmingly African American) ghettos in the United States, see Duneier (2016) and Shelby (2016). Different researchers define "ghetto" in different ways. Most commonly, these are neighborhoods that are predominantly black and where more than 40% of the population lives at or below the federal poverty line.

[2] My use of 'pathology' and 'deviance' is not meant to express any normative judgment, only to indicate that the phenomena so labeled are considered as such by "mainstream" (middle class, white) society.

Racial bias and discrimination, however, have always been central to the cycle of poverty framework. It's racism that supposedly first concentrates poor blacks in the ghetto. Not long ago this was of course a plausible mechanism. But racism at the level of explicit attitudes has considerably decreased since the 1960s, preventing it from explaining how the cycle of poverty hasn't just continued but *increased* and become more pernicious. If sociologists avail themselves only of the notion of explicit racism, the cycle of poverty framework won't even get off the ground in contemporary contexts, retaining historical interest at best. It seems time, then, to understand how *implicit* racial biases might factor into the cycle of poverty.

I turn to the most important conclusion first: Research suggests that racial discrimination is one of the major mechanisms that perpetuates entrenched poverty. Yet explicit racial bias has plausibly declined over a period during which entrenched poverty has persisted, if not worsened. I argue that implicit biases (against race, as well as particular socioeconomic classes and neighborhoods) fill this explanatory gap, and help us to make sense of sociological and economic patterns that would otherwise remain mysterious. As I argue in the next two sections, implicit biases are an integral part of the causes that currently perpetuate entrenched poverty in U.S. inner cities. In Sections 4 and 5, I argue that implicit attitudes also play an important role in the other arc of the cycle. Much of the negative impact of living in poverty is its cognitive effect on implicit attitudes, as Sendhil Mullainathan and Eldar Shafir (2013) have argued. In this way, implicit attitudes are among the causes but also the self-reinforcing effects of poverty. Many of the attitudes that hinder the ghetto poor from getting out of poverty and increase others' biases against them are implicit attitudes.

## 5.2 Concentrated Poverty

From early on in urban sociology, accounts of American inner city or "ghetto" poverty like Drake and Cayton's (1945) and Clark's (1965) have emphasized racial discrimination. In a decisive break from this tradition, William Julius Wilson has argued that while historical racism, primarily in the form of slavery and Jim Crow, played a role in *creating* ghetto poverty, the *maintenance* and increasing geographic concentration of said poverty during the 1970s and 1980s was much more a matter of economic factors than racism.[3] In particular, Wilson argues that the contemporary plight of the ghetto poor is primarily due to general economic restructuring and depression. Against Wilson, Douglas Massey and Nancy Denton (1993) present compelling evidence that economic restructuring alone cannot explain the increase in the concentration of poverty, and that racial discrimination is required to explain this continued residential segregation. In effect, Massey and Denton defend a version of the traditional vicious cycle theory—one which, I believe, is an improvement on Wilson's own account. However, I'll argue that without the notion of *implicit* bias, specifically, Massey and Denton's theory fails to address the very puzzle with which Wilson himself begins.

[3]Wilson does allow that institutional racism may persist as a holdover of earlier psychological forms of racial bias, but he downplays any ongoing effect of psychological racism in contemporary contexts.

In *The Declining Significance of Race* (1980/2012) and *The Truly Disadvantaged* (1987/2012), Wilson raises the puzzle of how inner city poverty could have become so much worse and so much more concentrated during the very period of the greatest Civil Rights reforms and decreased racism against black Americans. At the very least, decreasing racial hostility (starting in the 1960s and continuing steadily since) cannot explain a marked increase in black poverty (as happened in the 1970s–1980s). In ways I'll go on to question, Wilson (not unreasonably, writing before most of the psychological work on implicit bias had begun) assumes that racism cannot explain the perpetuation and increased concentration of poverty in inner city ghettos, and sets out to provide an alternative explanation.

In particular, Wilson (1987: 39–46, 100–104; cf. Wilson 1996) argues that the exodus of manufacturing out of the inner city, nationwide deindustrialization, and the switch to more of a service economy scuttled the traditional career bases of many poor young black men beginning in the 1970s. Blacks had come to rely more on manufacturing and other manual blue collar jobs more than other minority ethnic groups—partly the result of the Great Migrations and lack of other jobs they were educated for and were not barred from because of their race. Thus, the switch from a manufacturing economy to a white collar service economy created a "mismatch" between the skills most black workers had and those they were perceived to need for the jobs available. In turn, the lack of jobs led to a dearth of financially stable, "marriageable" men, increasing the prevalence of births out of wedlock and female-headed households—the primary recipients of government aid.[4]

At the same time, Wilson claims that Civil Rights legislation and affirmative action allowed middle-class blacks to move out of the inner city in record numbers for the first time, into neighborhoods they had once been unwelcome in. Whereas many inner city neighborhoods of the 1940s and 1950s had been thriving, socioeconomically-integrated communities—think the Harlem Renaissance of Alain Locke(1925)—the inner city of the 1970s and 1980s increasingly became one in which only the most destitute blacks were left behind. Because the services of *local* doctors, lawyers, teachers, priests, grocers and others from the middle class typically provide a crucial social buffer in weathering economic downturns, their departure from the community meant that deindustrialization hit the inner city that much harder. Wilson (1987: 56) argues that, in part, the flight of middle-class blacks removed traditional role models from the ghetto of the sort that children might learn mainstream values from.

These economic factors and patterns of emigration combined to lead not just to an increase in poverty for African Americans, but also to the increased *concentration* of poverty and its attendant social disadvantages within inner city ghettos, which became increasingly socially isolated from mainstream society. In turn, this concentration and isolation make their own, independent contribution to the problem, exacerbating it over and above the effect that the same level of average poverty would have on the same group spread out more sparsely into integrated neighborhoods. For example, it becomes more difficult to find jobs through informal social networks in neighborhoods with concentrated poverty that are socially isolated. And eventually, living in a neighborhood with few other role models leads

[4]This was especially the trajectory of the "rust belt" manufacturing cities of the Northeast and Midwest.

some black men to develop a hostile attitude toward menial work, increasing the mismatch between their skill sets and the preferences of potential employers.

The notion of concentration effects predates Wilson's work. Clifford Shaw and Henry McKay (1942) showed that cycles of crime and other delinquent behaviors in particular neighborhoods often persist over significant periods of time and are correlated with a whole host of other pathologies, like infant mortality, low birth weight, and other physical and mental health problems. In this way, concentrated poverty leads to *concentrated disadvantage* more generally.[5] Drake and Cayton (1945) went on to show that maps of incidence rates for health disorders, crime, and other social pathologies show striking overlaps; that "death and disease" in Chicago show particular geographic patterns.[6] Wilson's model, though, shows how structural-level, economic factors can interact with such concentration effects.

## 5.3 The Effect of Implicit Biases on Poverty

Wilson's model has all but defined dialogue in contemporary urban sociology. The primary objection to the framework has been that it inaccurately downplays the ongoing importance of racism. Douglas Massey and Nancy Denton (1993) preserve much of Wilson's emphasis on concentration and geographic factors, but argue that we cannot understand these without understanding racial discrimination not only as a historical variable, but as a persisting effect that continues to explain the maintenance of ghetto poverty.[7] According to Massey and Denton (1993: 118), the exodus of the black middle class alone is insufficient to explain the rise of concentrated disadvantage, as is economic restructuring on its own. Instead, these factors are better understood as exacerbating the necessary, and independently sufficient, effect of continuing racial discrimination—in particular, residential segregation.

Massey and Denton (1993) argue that ghettos are not the "natural" result of "like" preferring to live with "like" housing selection on the part of blacks (Schelling 1978), but are instead the result of racial discrimination operating through a number of channels: federal housing programs, national real estate organizations, court rulings, and also local factors including the practices of particular real estate agencies, realtors, and landlords, as well as restrictive covenants among homeowners in affluent neighborhoods. As blacks moved into Northeastern and Midwestern cities during Reconstruction (the first Great Migration) and after WWII (the second), their entry into white neighborhoods was contested through harassment, cross-burnings, bombings, and other overtly violent acts. These eventually gave way to less overt attempts to defend the "color line" (to keep blacks in the overwhelmingly black neighborhoods they already lived in). For instance, the use of restrictive covenants—

[5]This and other work in the "Chicago School" of urban sociology often parallels epidemiological research on the ecological concentration of disease (see, e.g., Goldberger, Wheeler, and Sydenstrycker 1920).

[6]For research showing that these same overlaps persist today, see especially Sampson (2012).

[7]For other work incorporating racism into Wilson's framework, see Massey (1990), Massey & Eggers (1990), Jargowsky (1997), Morenoff and Sampson (1997), Quillian (1999), and Massey & Sampson (2009).

agreements between homeowners not to sell to any blacks (or non-caucasians)—became increasingly common (even though declared unenforceable by the Supreme Court in 1948).

According to Massey and Denton (1993), racial discrimination also brought about segregation in the form of federal housing policy. The Home Owners' Loan Corporation (HOLC) and later Federal Housing Administration (FHA) institutionalized the practice of "redlining": these agencies used color-coded ratings systems to evaluate the risk of loans made for homes in specific neighborhoods, with the lowest, most risky category colored red. In part because of explicit racism, most poor black neighborhoods were thus red-lined and unlikely to receive loans. Not only did the HOLC and FHA fail to finance loans for homeownership in these areas, but private banks also used these agencies' ratings systems to determine their own lending practices, effectively shutting blacks out of much of the lending industry from the 1930s to 1960s (Massey and Denton 1993: 51–57). The Department of Housing and Urban Development (HUD), started in 1965, only made many matters worse, with plans for several decades based around high-rise housing projects that eventually came to be seen as a large mistake and, in some cases, were ruled to be discriminatory by federal courts.

The practices of individual real estate agencies and realtors contribute to the perpetuation of residential segregation, as well. Until 1950, the National Association of Real Estate Brokers' code of ethics contained an article stating that a realtor should never introduce "members of any race or nationality" into a neighborhood that would be "detrimental to property values in that neighborhood," a perspective that some maintain has persisted to the present day, albeit in a less overt form (Massey and Denton 1993: 37). Field studies have confirmed that real estate agents commonly engage in an array of discriminatory behaviors, ranging from flat-out refusal to show or sell properties to black families to various forms of subterfuge (e.g., claiming that a unit is already sold or unavailable, contrary to fact). And, if all else fails, whites tend to move out of neighborhoods in large numbers as soon as black immigration reaches a (very low) threshold, in which case the neighborhood eventually becomes an expansion of the ghetto that initial black immigrants were attempting to distance themselves from. More recently, Massey and colleagues (2013) have shown that various suburban zoning restrictions on population density also sustain residential segregation.

Perhaps most importantly, Massey and Denton (1993: 123–4) present the results of a number of mathematical simulations testing what effect different variables have on concentration effects. The main upshot of their findings is that while economic restructuring may have increased average black *poverty* since 1970, it cannot explain its increased *concentration* in particular areas, as this financial downturn would otherwise be diffused over relatively socioeconomically integrated neighborhoods. Instead, the models show that it was racial segregation that confined this increase in average black poverty to a small range of geographic areas. Hence, segregation, not deindustrialization or other structural economic factors, is primarily responsible for the cyclical "add on" effects of concentrated poverty and the concentrated disadvantage it begets. Overall, the simulations show that given a poor minority group, successively higher levels of racial segregation against this minority alone are able to explain increasing geographic concentration of poverty. As Massey and Denton (1993: 144) note, this is why dark-skinned Puerto Ricans are the only Latinos to experience

concentrated poverty on the order of African Americans. While they're otherwise identical to lighter-skinned Puerto Ricans—who fare much better in the United States—racial bias based on skin color consigns dark-skinned Puerto Ricans to a similar fate as other blacks.

Much of Massey and Denton's amendment to Wilson seems an improvement. But for the reasons Wilson had already pointed out, assigning explicit racism too large an explanatory role leaves the cycle of poverty framework increasingly inapplicable. At the very least, the framework would fail to explain how rates of concentrated poverty and disadvantage increased while levels of explicit racism were dropping. Massey and Denton's picture, without supplement, would not explain how discrimination on the part of realtors, government officials, white homeowners, and others would have led to increased segregation, since the System 2 explicit racist attitudes that support these forms of discrimination were subsiding. By this point, it should be clear how we can easily supplement Massey and Denton's picture, however. While levels of explicit racism dropped starting in the 1960s, rates of *implicit* racism may have remained level all the while. Even if realtors, government officials, and white homeowners are explicitly committed to egalitarianism (which many are), implicit racism would explain why they nonetheless systematically prevent blacks from leaving the ghetto. As the research of Chapter 2 showed, statistically speaking we can be confident that many of these agents do harbor implicit racial biases that do influence their actions.[8] In that case, implicit biases would explain why economic downturns—even in the absence of explicit discrimination—would still be filtered into particular geographic areas, leading to increases in concentrated poverty and disadvantage. If the poor become poorer and are forced into the inner city, then the inner city becomes more poor. The proximate psychological mechanisms of much institutional racism are likely realized by implicit biases, as well. In short, we can fill the glaring hole in the cycle of poverty framework that Wilson has drawn so much attention to with implicit bias. If, as Massey and Denton plausibly argue, persistent racism plays a crucial role in the perpetuation of concentrated poverty, we can expect it to operate at the present time in large part through the now-familiar mechanisms of implicit bias.

Some urban sociologists have recently shown signs of appreciating just this. In *Great American City* (2012), Robert Sampson develops a cycle of poverty model that draws heavily from Wilson's insights, but also assigns a large role to implicit bias. According to one longstanding sociological current, culminating in the "broken windows" theory of the 1990s, ghettos primarily remain poor because they're highly *disordered*—that is, they lack the capacity for collective community action. Public disregard for norms (like the appearance of broken windows) in a neighborhood leads those in and outside of the neighborhood to infer that it's easier to commit crimes there without repercussion, thereby increasing such crime. Sampson shows that, in fact, the more important predictor of violent crime rates is *perceived disorder*, rather than observable, objective signs of disorder in the actual environment (litter, graffiti, vacant housing, public drinking, fighting, drug dealing, and the like). According

[8]Much of this residential discrimination likely occurs due to hijacking of the sort outlined in Chs. 2 and 3, and so often goes unnoticed by its perpetrators, operating as it does by subtly tipping the weight of different attributes in complex decisions. This doesn't necessarily excuse its perpetrators, however, since much of it's likely motivated by implicit biases that also count as ISAs of the sort outlined in Ch. 4.

to Sampson (2012: 129–131), the psychological mechanisms that underlie such perceived disorder and a neighborhood's "reputation" are implicit biases (or other implicit attitudes).

Sampson (2012) compares levels of perceived disorder gleaned from the community survey component of the Project on Human Development in Chicago Neighborhoods (PHDCN) with comprehensive levels of objective disorder coded from videos of the census block groups where respondents lived. Concentration of black and minority residents is a better predictor of perceived disorder than objective disorder, confirming the implication of implicit bias. Perceived disorder (in 2002) is also a better predictor of the homicide rate (in 2002-6) than objective disorder. And earlier perceived disorder (in 1995) significantly predicts later homicide rates but "*concurrently observed disorder does not*" (Sampson 2012: 147). Understood against the backdrop of Massey and Denton's contention that racial discrimination is responsible for the concentration of poverty and disadvantage, these results strongly suggest that the implicit biases of those who live outside the ghetto—i.e., of "mainstream" Americans—are influencing the future delinquent behavior of those who live within. Sampson suggests this effect is mediated by part of the cycle of poverty that constitutes a self-fulfilling prophecy: if the urban poor are aware of the social perceptions of mainstream society, they may come to see their own neighborhood as disordered, weakening their motivation or resolve to combat objective disorder. That, in turn, may well increase objective disorder.

Importantly, the implicit attitudes that underlie perceived disorder are not just racial biases. They also include implicit biases against particular neighborhoods and classes—associations, for example, between GHETTO and POOR and VIOLENT or IRRESPONSIBLE. As Tommie Shelby (2016: 47) notes, the ghetto poor must face the interaction of implicit biases against race, class, and place. Kirschenman & Neckerman (1991), for instance, find that after controlling for racial and other biases, biases against particular neighborhoods still affect employment decisions, presumably because residence in these neighborhoods is implicitly associated with things like violence and a poor work ethic.[9] Indeed, this connection between different associations also seems to be what explains a point made by Massey and Denton (1993: 160)—that concentration effects facilitate increased racial discrimination. This is how the cycle of poverty comes full circle. As Mitchell Duneier (2016: 223) puts it:

> When schools or streets or hospitals are rendered unequal through societal power, they come to symbolize the black way of life. This way of life is made visible through the physical living space that becomes known as the ghetto. Once the ghetto can be apprehended as a physical reality, subjective perception plays a major role in its perpetuation: the association between blacks and the observed physical conditions becomes a rationalization for further discrimination."[10]

Some of the best examples of how perceived disorder can lead to objective disorder are illustrative. It was easier for New York City to reduce its number of fire departments in

[9]Social psychologists take note: Very little experimental laboratory work has been done on implicit associations with socioeconomic class let alone place, especially compared to the volume of work on race.

[10]Duneier (2016: 161) further suggests that spatial isolation produces delinquent behaviors which are then "stigmatized, seen as innate, and used as evidence to justify continued spatial isolation."

specifically black ghettos in the 1970s, for instance, given that it was easy to pinpoint the areas in which poor blacks with little political power lived. The epidemic of fires this produced increased objective disorder. More generally, geographic concentration makes it easier for people with biased attitudes to sap funds from hospitals, schools, police and fire departments, and other public institutions in biased ways (Myrdal 1944: 618). Indeed, associations between race, class, and place make it that much easier to rationalize discrimination based on class and race as instead being motivated by a concern about place (that particular neighborhoods are dangerous, for instance). Violent crime, likewise partly caused by racial segregation-produced disadvantage, becomes a justification for massive rates of incarceration, further isolating the ghetto poor by extending the cycle of poverty through prison walls.

Massey and Denton (1993: 162–5) provide another striking example of how perceptions of the ghetto poor can affect them in ways that subsequently worsen mainstream discrimination against them: the rise of Black English Vernacular. Research by the linguist William Labov (1972, 2012) suggests that while African Americans spoke a different dialect of English during slavery and Reconstruction, it wasn't until several decades after WWII that Black English Vernacular developed, suggesting that it's the direct result of racial discrimination-based neighborhood segregation.[11] In turn, speaking Black English Vernacular is seen as inferior by members of mainstream society and constitutes a serious handicap in educational and occupational contexts, ultimately leading to even more stigmatization of ghetto residents.

In sum, Wilson is surely right that deindustrialization and other economic factors contribute to the perpetuation of ghetto poverty. He's also right that explicit racism cannot explain the rise in black poverty in the 1970s–1980s. Massey and Denton are nonetheless correct that economic downturns are mediated by discrimination-based segregation—so long as we understand this as implicit discrimination—and that implicit discrimination of this kind contributes significantly to the concentration of poverty in certain areas and the "ratcheting up" of the whole cycle of poverty in the ways carefully brought out by Wilson. Implicit biases are also well-suited to explain why increased concentration of poverty and disadvantage in the ghetto might only make attitudes toward blacks more negative. We might not expect this—so much as sympathy—on the part of most Americans' explicit attitudes about race. But their connections with place and class make implicit racial biases directly, automatically associated with the further (increasingly negative) properties encoded by or tied to these other concepts—DISORDERED, VIOLENT, and the like. Indeed, given what we know about how implicit associations work from earlier chapters, we might predict that increases in the concentration of poverty, disadvantage, and the behaviors they engender would precisely lead to *increased* implicit bias against blacks, further increasing segregation. Implicit biases—not only to race, but also socioeconomic class and place—are thus crucial to understanding how and why the cycle of poverty continues in U.S. ghettos.

[11] Blacks and whites were not geographically segregated in the South at comparable rates during slavery.

## 5.4 The Effect of Poverty on Implicit Attitudes

A great deal of research shows that implicit biases tend to be internalized by those they're biases against, albeit not as strongly as they're internalized by members of the out-group. In some of the pioneering work in this area, Kenneth Clark and Mamie Clark (1941/1958) showed that even black children prefer to play with white rather than black dolls. There's also been extensive (though controversial) research on "stereotype threat." According to Claude Steele (2010), implicit biases lead to a type of self-fulfilling prophecy generally. Being reminded of their gender, for instance, leads women to underperform on mathematical and other tests associated with masculine skill-sets. Here, I want to focus on the implicit effects of poverty, as such. Many theorists have claimed that poverty's effect on the poor is exactly part of what "ratchets up" levels of implicit bias against the poor, making it a crucial mechanism to understanding the second, "return" arc in the cycle of poverty.

While there are other effects of poverty both on the external environment and the cognition of the poor (see Shelby 2016: 120 for its effects on the mental health of children), Sendhil Mullainathan and Eldar Shafir (2013) have recently collected an impressive body of evidence suggesting that scarcity of resources, as such—independent of race and other typical correlates—creates its own "mindset."[12] In particular, Mullainathan and Shafir (2013: 44–5) argue that scarcity (of any resource, but money is a common currency) captures implicit attention, such that one has trouble thinking of anything else. Starvation makes it difficult to think of anything but food, the lonely think of little else than companionship, and poor American children even see US coins as larger than rich children (Bruner and Goodman 1947, Saugstad and Schioldborg 1966). Scarcity puts one in an implicit mindset that makes one more attentive to and efficient with respect to the scarce resource, but also creates a type of tunnel vision, causing one to neglect potentially more important things. Mullainathan and Shafir (2013: 29) argue that this type of "tunneling" can (i) make people fail to consider certain options in the first place and (ii) affect their cost-benefit and other practical reasoning. By focusing so much attention on the resource that's scarce, consideration of and deliberation about all other resources is inhibited—again, all on the implicit level.

Scarcity has general effects on one's mindset. Mullainathan and Shafir (2013) show that scarcity overtaxes our capacities to compute, pay attention, make good decisions, stick with plans, and resist temptation—in short, it taxes our general cognitive capacity and "executive control." In one experiment, people in a New Jersey mall were given measures of IQ and fluid intelligence (Raven's Progressive Matrices) and measures of executive control (tasks that require participants to inhibit their automatic, impulsive responses). What they found was that prompting subjects to think about serious financial decisions significantly reduced scores on the Raven's matrices and impulsive control tasks, but only for poor subjects—not for wealthier participants. Thus, the poor seem to have a standing dispositional propensity to be hijacked by the scarcity mindset. When they're reminded of the source of scarcity in question, their general capacities for self-control become overtaxed. In contexts of concen-

[12]See also Shah, Mullainathan, and Shafir (2012) and Mani, Mullainathan, Shafir, and Zhao (2013).

trated poverty, this propensity for being hijacked is surely all the more frequently activated, and Mullainathan, Shafir, and colleagues replicate these same effects in the field with a large set of Indian sugarcane farmers, who show lower fluid intelligence and executive control scores when money is more scarce (pre-harvest) than when it's not (post-harvest).

By presenting constant internal distractions that draw one's train of thought ever back to the source of scarcity (food for the dieter, a jackpot for the gambler), scarcity depletes how much willpower one has to deploy in general. According to theories of *ego depletion*, willpower is a fixed resource, such that overtaxing it leaves less willpower to use for other purposes (Baumeister et al. 1998). Because depletion only makes it harder to resist further temptation, scarcity has self-reinforcing effects. Indeed, even more basically, the scarcity mindset leads to a self-reinforcing vicious cycle because it leads the poor to over-borrow (Mullainathan and Shafir 2013: 115). In this way, a person can fall into a "scarcity trap" where their behavior compounds their own scarcity. In another lab-based study, Mullainathan, Shafir, and colleagues asked subjects to play a video game based off of Angry Birds, *Angry Blueberries*, in which players slingshot blueberries at waffles for points. "Poor" subjects are given 3 blueberries per level, and "rich" subjects 15 shots per level (with 10 levels total). In some conditions, subjects were able to borrow shots from future levels (with 100% interest), whereas in other conditions no borrowing was allowed. Blueberry-poor subjects did better on each average shot and earned more points per shot than the blueberry-rich. (Scarcity focuses one's attention on the source of scarcity.) However, the blueberry-poor subjects who could borrow shots from future rounds also earned fewer points than poor subjects who could not borrow. The poor subjects are tunneled in by the scarcity of shots on the current round and neglect attention to future rounds, such that when the option is available, they can easily be induced to over-borrow resources from future rounds, ultimately earning fewer total points.[13] This is striking evidence of how scarcity of even an arbitrary resource, independent of other factors, drives a vicious cycle in the minds of the poor: as Mullainathan and Shafir argue, it creates a mindset conducive to thinking and behavior that are not only more likely to keep one in, but to *deepen*, one's poverty. These patterns of thinking and behavior increase actual poverty, which then exacerbates the scarcity mindset, ratcheting up the cycle.

David Harding (2010) presents comparative interview data showing that hijacking affects the reasoning of adolescent boys from poor neighborhoods more than matched boys from less poor (working class) neighborhoods. Harding finds that growing up in poverty affects many of the boys' (perceived) option sets in damaging ways—e.g., regarding educational and career decisions. Poverty completely blinds boys to some options that are in fact open, and it also blinds them to the incompatibility or tension between the means toward different ends—e.g., that pursuing vocational and technical classes in high school may not bolster applications to four year colleges. More surprisingly, boys from poor neighborhoods are more likely to see some options as open that are not, in fact, genuinely available (or to overestimate the likelihood of their occurrence). Harding (2010: 209–211) finds that boys from disadvantaged

[13]Poor subjects who can borrow with interest also do worse than those who can borrow but without interest. Rich subjects do similarly across all conditions (since they do not borrow at comparable rates).

neighborhoods persist in dreams of becoming professional athletes and famous musicians significantly longer than otherwise-matched boys from slightly more affluent neighborhoods. That is, it takes poor boys longer to realize that stardom is an extremely unlikely goal. These results corroborate that hijacking of the general sort Mullainathan and Shafir (2013) provide lab-based evidence for can and does affect the actual reasoning of ghetto residents.

Situating this work against the backdrop of the last section: the behavioral patterns that the scarcity mindset and persistent hijacking lead to are observed from afar by members of mainstream society. Lack of intelligence, willpower, inhibition, and irrational financial habits come to be implicitly (if not explicitly) associated with the poor, and hence in many contexts, come to be associated with blackness. Even when people are being carefully explicitly egalitarian, it can be very hard to prevent implicit associations from encoding morally unfortunate correlations in the environment. As Tamar Gendler (2011) stresses, implicit associations arise from our natural tendency to categorize, which is difficult to avoid. For instance, it may be very hard not to implicitly associate the concepts BLACK and VIOLENT if there is, in the (unjust) world as one experiences it, in fact a correlation between blackness and violence. Sarah-Jane Leslie (forthcoming) argues that many implicit biases arise from our primitive, default method of generalization (which is best expressed using generics). Because this mechanism is especially sensitive to dangerous properties of individuals, it tends to generalize these, in particular, to the rest of the group those individuals are members of.

## 5.5 The Effect of Poverty on ISAs

In good ethnographic work, one can see the interpersonal, intergroup cycles between mainstream society and ghetto society play out in real-time. Children raised in the ghetto glean most of their exposure to the outside world from the media, which paints an unmistakably inferior picture of their lives. Sudhir Venkatesh (2000: 43–4) reports that residents of Chicago's Robert Taylor Homes (the biggest low-income housing projects in the country before they were razed to the ground) were acutely aware that they lived in a uniquely bad environment—exposed, as they were, to popular music, film, and even newspaper stories going so far as to compare living in the Robert Taylor Homes to "living in hell" (Walinsky 1987). As geographic isolation leads to social isolation, the only direct contact many inner city children are likely to have with mainstream society is through school, social services, work, and the criminal justice system, where the biases of their case workers, employers, and the police is often all too apparent. Even the schools, according to many, are merely a staging ground for street life, with teachers and school officials—partly out of necessity—taking on the role of guards and wardens (Wilson 1987/2012, Anderson 1999, Venkatesh 2000).[14]

[14]Venkatesh also stresses the important role of ecological factors other than (differences between) neighborhoods. Venkatesh (2000: 8) notes, for instance, the "visibly jarring" and imposing architectural presence of the Robert Taylor Homes themselves over the surrounding landscape. The design of public housing high-rises (with their attendant problems with elevators and stairwells) and the fact that such buildings often occupy only a small area of the land on which they're located further contribute to concentrated disadvantage.

Controversially, the combined effects of living in poverty and exposure to mainstream society and its biases (however indirect) may lead to unique cultural traits within the inner city. Some sociologists implicate a "culture of poverty" or "subculture of scarcity" in many of the delinquent behaviors prevalent in the ghetto (Lewis 1961, 1966, 1968, Horowitz 1983, Sánchez-Jankowski 1991, 2008). And there's growing recognition that culture is largely an implicit affair. Stephen Vaisey (2009) argues that culture, as thought of by sociologists, is predominantly subserved by implicit attitudes, and he shows that possession of different implicit moral cultural frameworks has significant effects on individual delinquent behavior (for instance, cheating on tests and drug use).[15] Presumably, implicit moral frameworks and other aspects of culture are not subserved by "mere" implicit attitudes, but by implicit self-associations (ISAs), in particular, as outlined in Chapter 4. Recall that ISAs are hierarchically organized and often mediated by associations with one's in-group (Sechrist & Stengor 2001, Otten 2003), where in-group likely anchors one of the strongest self-associations for most individuals. Living in poverty and being exposed to others' implicit biases may thus have effects on what the poor implicitly value and care about. Hence, if there are unique cultural traits involved in the cycle of poverty, these are plausibly realized by ISAs.

Elijah Anderson (1990, 1999), for instance, discusses how many inner city youth adopt a *code of the street*—an informal set of cultural norms that gives pride of place to "respect" gained through physical violence (and overt displays of wealth).[16] Often, following the code of the street means being prepared to use violence (or to at least feign willingness to do so) in order to physically protect oneself. For those on the street, adopting the code can be a matter of survival. And once one learns the appropriate roles and scripts, one may find oneself carrying them out implicitly. Even those who are not willing to use violent force to protect themselves often posture as though they are, making it that much harder

[15] See also Hitlin and Vaisey (2013) and research from the "Measuring Morality" project.

[16] Venkatesh (2000, 2008) expands this to a code of "hustling" (any illicit economic activity) generally. The "code of the street" and subcultures of poverty are in many respects "honor cultures," making them one instance of a more general, ecologically-driven pattern. *Honor cultures* are characterized by a willingness to use violence to protect one's reputation, and arise whenever individuals are at economic risk from others and the local state is too weak to protect against theft (Nisbett and Cohen 1996). This is especially true when property is highly portable, as with livestock compared to field crops and drugs and illicit merchandise compared to other forms of capital. When your means of financial well-being could easily be stolen and you're the only one who might prevent this, it makes sense to have a reputation of using violence to do so. Nisbett, Cohen, and others provide evidence that the herding-based honor culture of the Scots-Irish migrated with them to Appalachia and still persists in the Southern compared to Northern United States (e.g., in the form of higher cortisol levels in response to aggression and greater willingness to condone the use of physical violence to preserve one's reputation). This held-over honor culture likely expanded westward with "frontier culture," and with poor whites as they moved out of the South into areas of urban manufacturing in Northern cities. Thomas Sowell (2005) argues that some ghetto culture is another remnant of this Scots-Irish honor culture, and it's not implausible that urban African Americans—as a result of their residential proximity to poor rather than affluent members of other ethnic groups—only had these cultural effects reinforced after immigrating to Northern cities themselves. In sum, the origins of the "code of the street" and "culture of poverty" may have much less to do with blackness or the unique plight of inner city African Americans, per se, so much as with certain ecological conditions that can arise and be sustained through various realizers.

for already-biased members of mainstream society to distinguish between those who truly value, e.g., violence, and those who merely put on to. The code also reinforces a resistance to menial, dead-end work, leading many employers to adopt the policy of not hiring anyone who resides within inner city zip codes at all. Matthew Desmond (2016) has recently shown how these types of mutually reinforcing forms of intergroup confirmation bias play out in the Milwaukee housing market—landlords who own low-income housing often unable to turn a profit if they keep these properties well-maintained; low-income tenants often coming to see said landlords as biased and exploitative, causing them to further neglect these properties.

As differences in values between mainstream and "street" culture widen, teachers, case workers, employers, and the police perceive their biases being justified and these in turn increase in strength, ratcheting up the vicious cycle of poverty. Again, this process of strengthening implicit biases need not occur via System 2 mediation—it can be driven simply by changes in strength of correlation within the environment (Gendler 2011, Leslie forthcoming). Once one associates GHETTO and VIOLENCE, increases of violence in the ghetto will strengthen the GHETTO/VIOLENCE association, even in the absence of any explicit, conscious thought. In other words, the whole cycle of poverty can be perpetuated, at each step, without anyone's explicit endorsement, intention, or even awareness. Concentrated poverty and implicit biases may lead to cultural and other implicit effects on the minds of the poor, and these may in turn compound the level of implicit bias against them.

## 5.6 Conclusion: Moral Ecology

Poverty in U.S. ghettos must be understood as an intergenerational phenomenon—something passed down from parents to children that plays out over "*a series of* vicious cycles... in a spatial context" (Duneier 2016: 227; cf. Sharkey 2013). Earlier cycles have involved high levels of explicit racism, to be sure, and some level of explicit racism no doubt remains. However, we've reached a point in the series where much of each cycle—both its forward and return arc—is implicit, rather than explicit. Recognizing this role of implicit attitudes is crucial for the continued viability of the cycle of poverty framework, down to its foundations. My hope in this chapter has been to explain what drives cycles of poverty today—and how they're still able to persist even *if* all parties have only the best of explicit intentions.

It's now clear how, in social combination, implicit biases can have serious effects on others' autonomy. Implicit biases against race, class, and place—both on the part of mainstream society and many poor blacks themselves—seriously constrain poor blacks' autonomy. Hijacking is part of the effect on autonomy, but understanding the construction and maintenance of inner city ghettos is also crucial, as geographic concentration effects compound or "ratchet up" the impact of hijacking.[17] The importance of spatial concentration in perpetuating cycles of poverty makes the foregoing discussion a posterchild for moral ecology of the sort introduced in Chapter 1.[18] Here, the focus has been on how implicit attitudes—in

[17]The existence of U.S. ghettos surely has other effects on the autonomy of poor blacks, as well.

[18]For a related (but overtly normative) use of the term "moral ecology," see Vargas (2013a: 244–245).

particular, implicit biases—operate in the wild. The discussion is relevant to moral psychology in at least three ways. First, it shows how phenomena of huge moral importance—like poverty—are partly ecological phenomena. Second, it shows that we cannot truly understand implicit biases themselves without a proper understanding of how they fit into their moral landscape. The psychological is partly ecological. Third, our discussion shows that to intervene responsibly on many implicit (and other) attitudes, we must appreciate the ecological context in which they operate. To understand how poverty and implicit biases work in any way that we can put to much work, it's necessary to go beyond the lab.

Philosophers working in the *situationist* tradition have argued that features of context can affect moral behavior and autonomy (Doris 2002, Harman 1999). But as noted in Chapter 1, situationists have focused on whether small, surprising, seemingly morally-irrelevant features of situations can have such effects. For example, Isen and Levin (1972) found that far more people who had just found a dime in a phone booth helped a passerby who dropped a stack of papers compared to people who had not just found a dime. Unfortunately, far less attention has been paid to how large, obvious features of situations affect moral behavior and autonomy.[19] If whether or not one just found a dime has such an enormous impact on helping behavior, imagine the effect of whether or not one grew up in the ghetto. Moral ecology goes beyond these relatively one-off, unsystematic lab findings and attempts to build a more complete (albeit messier) but still scientifically well-backed model of contextual effects.[20]

Turning to the first conclusion, one of the major findings of urban sociology is the importance of context and place in understanding poverty and the delinquent behaviors that often attend it. Researchers have consistently found strong correlations between the locations of "death and disease" within cities for decades (Drake and Cayton 1945), and the presence of *neighborhood effects* is now well-documented: even after controlling for average income, race, and numerous co-occurring factors, what neighborhood one lives in still has independent effects on a number of morally-relevant variables, including crime rates, helping behavior, and economic success (Wilson 1987, Massey and Denton 1993, and Sampson 2012). Insofar as we want to understand poverty and its effects, then, we have an incomplete picture unless we look to ecological factors like neighborhood.[21] More basically, we cannot understand cycles

[19]That said, Stanley Milgram (1970) of the "obedience to authority" experiments fame himself discussed the relevance of living in a city to helping and harming behavior (which he thought exacerbated the bystander effect), and Nisbett and Cohen's (1996) work on honor cultures is a classic of situationist social psychology.

[20]Philosophers have also noted that situationist results may present problems for many epistemological theories, not just moral theories and theories of agency (Olin and Doris 2014). Continuing the current thread, then, taking large, obvious features of context seriously may suggest the need for an "epistemic ecology." (Much of this research project would no doubt overlap with and be informed by "social epistemology.")

[21]Indeed, this overcomes one of the most serious shortcomings of situationism: Even if the laboratory studies show that internal character traits are not robust across situations, they don't show that the situational interventions used (e.g., mood, dimes, and the like) are situationally robust, either (Funder & Ozer 1983). Often, laboratory studies precisely employ seemingly trivial aspects of the situation that don't share any obvious connection to one another, and are relatively rare. In stark contrast, the effects of concentrated disadvantage on behavior are likely enormous, and urban sociology already provides considerable guidance in understanding why the factors comprising disadvantage are internally correlated in the ways they are.

of poverty without understanding *concentration* effects, an inherently ecological notion.

Second, we cannot do good social psychology without doing moral ecology. In particular, we cannot understand a fundamental aspect of how implicit biases against place, race, and socioeconomic class work without understanding the ecological mechanisms that tend to lead to their gradual self-reinforcement.[22] If implicit biases didn't tend to produce residential and other forms discrimination, and discrimination didn't tend to increase concentrations of poverty and disadvantage through segregation—inherently ecological and geographical factors—implicit biases would not reinforce themselves in the same largely automatic way that they're observed to. As Massey and Denton (1993), Duneier (2016), and Shelby (2016) note, the specifically geographical concentration of poor blacks in particular neighborhoods puts them at increased disadvantage, beyond poverty alone. My further claim here is that, for the reasons stressed by Gendler (2011) and Leslie (forthcoming), this geographic concentration also reinforces and strengthens implicit biases against place, race, and class. This self-reinforcement through discrimination-based concentration is part of the "ecological profile" of these implicit attitudes—how they work "on the ground." We cannot understand this fundamental aspect of the nature of implicit biases without doing moral ecology.[23]

Third, and perhaps least controversially, we cannot understand topics like poverty or implicit bias in the way needed to make responsible, informed interventions on them without doing moral ecology. Moral ecology suggests that laboratory psychology should only be applied to actual public policy with caution, and only under the supervision of an appreciation of the larger social context. Certain interventions or "nudges" might be appropriate in some ecological contexts, but not others. In the next chapter, we turn to such matters of policy.

[22]We also cannot fully understand how implicit biases based on race function in the real world without understanding how these tend to be associated with implicit biases based on socioeconomic class and place.

[23]I've focused on implicit biases in the context of U.S. inner cities, but the phenomenon is more widespread. Similar interactions between implicit biases against class, place, some racial or ethnic minority and concentration effects play out in the favelas of Rio De Janeiro, the slums of Calcutta and Nairobi, the barrios of Bogotá and Mexico City, the banlieues of Paris, council estates of Glasgow and London, and the townships or ikasi of South Africa (and to some extent isolated rural areas of the U.S. like Harlan County, Kentucky).

# Chapter 6

# Hijacking the Hijacking of Reason

## 6.1 Introduction

There are far more Americans who say they're willing to become organ donors than those who actually follow through and become organ donors (Kurtz & Saks 1996). However, in countries where you have to check the box on the back of your driver's license in order to opt *out* of organ donation, rather than *in*, far more people donate their organs (Johnston and Goldstein 2003). Knowing this, should the federal government have U.S. states that still have an "opt in" system change their default, increasing the number of donors?

Suppose the honor system for your office coffee fund isn't working: people drink plenty of coffee, but there's never nearly enough money in the donation tin to buy coffee supplies. You learn how Bateson, Nettle, and Roberts (2006) increased contributions to their office coffee fund *threefold* when they put an image of eyes (rather than flowers) above the donation tin. Apparently, being "watched" increases how helpful one's implicit, System 1 attitudes induce one to be (perhaps through associations between the presence of conspecifics and one's reputation), even when the eyes are not photographs and are merely schematic line drawings. Knowing this, should you put a picture of eyes on *your* office coffee fund? The *watching eyes effect* also makes people more morally well-behaved in a number of other situations.[1] So where should we stop? Should the government start posting billboards of big googly eyes on playgrounds? Above the stock exchange? Violent inner-city street corners?

These are the types of question raised by one of the most widely discussed academic books of recent years: Richard Thaler and Cass Sunstein's (2008) *Nudge*.[2] Much of the intrigue surrounds Thaler and Sunstein's advocacy of "libertarian paternalism," the proposal to use dual-processing research to design public policy—"nudging" people to make better decisions by influencing their implicit attitudes, specifically. Both Barack Obama and David Cameron

[1] Exposure to pictures of eyes has been shown, for example, to increase generosity and offers in the dictator game (Haley & Fessler 2005; cf. Rigdon et al. 2009), to increase donations in a public goods game (Burnham & Hare 2007), and to decrease littering in a self-serve cafeteria (Ernest-Jones et al. 2011).

[2] See also Thaler & Sunstein (2003a, 2003b, 2006) and Sunstein (2013, 2014, 2015, forthcoming).

established "Nudge Units," and Sunstein himself was the first head of the White House Office of Information and Regulatory Affairs. Not everyone has shown the same enthusiasm.

Most critics of libertarian paternalism have argued that engineering people's "choice architecture" to nudge them amounts to manipulation and violates citizens' autonomy. I argue that these worries are largely unfounded. As past chapters have demonstrated, we are often "manipulated" by our own implicit attitudes and the situational features that influence them already. We are thus "forced to nudge"—even when our eventual choice is to leave the influences on implicit attitudes that already exist in place, undisturbed. However, the complexities of past chapters also show that for many of the most important policy interventions, the normative question about how to nudge can only be settled with the aid of a more comprehensive and cross-disciplinary analysis of the situations they would change. To nudge responsibly, we need not only psychology and behavioral economics, but also sociology. This is no (general) objection to nudges, but it is a call for caution and humility.

In section 2, I introduce the notions of nudges and libertarian paternalism. In section 3, I discuss and rebut the objection that nudging, as such or in general, undermines autonomy in any problematic way. Nonetheless, the permissibility of nudges must be assessed on a case-by-case basis, and here, I focus on the types of implicit attitudes discussed in the last chapter. Section 4 thus corresponds to section 3 of the last chapter, and section 5 corresponds to sections 4 and 5 of the last chapter. In section 4, I discuss attempts to nudge implicit racial biases of the sort that perpetuate poverty through residential (and other forms of) discrimination, and in section 5 I turn to the topic of nudging the implicit attitudes of the poor themselves. I argue that while nudges targeting the poor are in many cases permissible, nudging people's implicit self-associations (ISAs) must meet a higher justificatory bar than nudging mere "first-order," non-SELF-involving associations (see Chapter 4), since nudging ISAs raises additional concerns about autonomy. Especially in these cases, to determine *which* way we *should* nudge, we need to understand the myriad causal influences involved from multiple levels of analysis, and so need to look to research from multiple scientific fields. To nudge responsibly and with high impact, that is, we need moral ecology. My project, as it has been throughout, is not to answer the normative question of whether to use any particular nudge (though when we are forced to nudge, it may be that nudging in *some* way—perhaps by omission—must be permissible). Instead, I mean to provide the bridges between disciplines that those who wish to pursue the normative questions will need.

## 6.2 Nudges and Libertarian Paternalism

Nudges are mainly defined by example, but involve influencing implicit attitudes to alter how people see the choice sets they're confronted with without removing any options in those sets or making any overly costly. Thaler and Sunstein (2008: 5-6) define a *nudge* as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives," where *choice architecture* is the organization of "the context in which people make decisions." For

example, the manager of a cafeteria or supermarket must choose how to display the food and other products, in what order and layout, and the like, and these differences in contextual framing will influence customers' preferences between options in ways discussed in Chapters 2 and 3. Recall that people prefer products placed on the right of a display, even when all the products are identical (Nisbett and Wilson 1977). Nudges are not meant to interfere with people's ability to choose as they prefer—they're libertarian (liberty-preserving). And nudges are meant to be paternalistic only insofar as they genuinely benefit those nudged. As Thaler and Sunstein (2008: 5) intend it to be understood, *libertarian paternalism* advocates only those nudges that "will make the choosers better off, *as judged by themselves*."

Nudges are typically intended to correct for "inherent" forms of implicit interference unearthed by the dual-processing research—instances of beating, bypassing, hijacking, and heuristics and biases—*by* influencing implicit attitudes in some way that counteracts that inherent interference, and subsequently changes behavior. In essence, nudges are aimed at System 1—they're not appeals to one's reasoning and explicit, System 2 deliberation, and they don't provide reasons or incentives. It's in this sense that nudges are more suspicious than mere advice, public information campaigns, or warning signs (at least in terms of their informational content). As Hausman and Welch (2010) and Reiss (2013: 293) note, nudges crucially do *not* involve rational persuasion. Seat belt fines, for instance, are not nudges.

Other examples abound. Southern California Edison gives customers "Ambient Orbs" which glow red instead of green when consumers use more energy, reducing energy use by 40% (Thompson 2007), and DIY Kyoto sells Wattson, a display of your energy use that can be uploaded to the internet and compared to other users'. These devices are so effective because they make energy use *visible* to users (in a nonintrusive way), following a common theme in nudge design. City officials have made the surface lines on Chicago's Lake Shore Drive narrower along its "S curves," for instance, making it seem like driving speed is increasing around these curves, causing drivers to slow down (Thaler & Sunstein 2008: 37–39).

Thaler and Sunstein focus on nudges by the government and private organizations or employers, but there are also self-nudges. For example, Christmas savings clubs (which you contribute money to over the year but can't withdraw any money from until just before Christmas), programs that automatically donate your money to your least favorite charity if you fail to achieve some important self-set goal (e.g., quitting smoking), gambling self-bans (which addicts can sign up for in order to have themselves barred from casino floors), no-bite nail polish, and disulfram all count as ways in which one can nudge oneself. Some self-nudges, like smart phone apps, have seemed potentially objectionable (in part because of their association with B. F. Skinner and cognitive behavioral therapy; see Freedman 2012 and Murray and Lombrozo 2017), but it seems clear that the majority of self-nudges are autonomy-preserving or -enhancing measures, not autonomy-undermining. In any event, most self-nudges pose no more threat to autonomy than whatever Ulysses loses when he ties himself to the mast to avoid steering toward the Sirens, and most opponents' objections instead center on nudging by the government and actors other than those who are nudged.

The nudges that have received most attention are those that would change economic and financial choice architecture. These include programs to help people invest and save in ways

that overcome various implicit heuristics and biases—often, overcoming the temptation of immediate gratification by changing the default or having "automatic" enrollment plans. The Save More Tomorrow plan, for instance, automatically increases the amount of one's salary put into savings every time one gets a raise (unless one actively chooses the non-default option of opting out of the increase), which doesn't make the increased contribution feel like a loss (Thaler and Sunstein 2008: Ch. 6. Recall the *loss aversion* effect discussed in Chapter 3). Indeed, automatic enrollment savings plans are among the most prominent nudges that have recently been promoted by the federal government (Gale et al. 2009).

Some purported nudges are harder to classify and more controversial, like mandatory calorie labels at chain restaurants and graphic warnings on cigarette packages (Sunstein ms.). Thaler and Sunstein's (2008) stable of nudges officially includes warnings and informational campaigns. But these don't involve the element of coercion that's usually thought to be essential for paternalism, at least insofar as these warnings and campaigns merely provide information (or incentives) that appeal to one's explicit, rather than implicit, attitudes. According to Daniel Hausman and Brynn Welch (2010: 127-128), for something to count as paternalistic rather than "rational persuasion," it must involve an element of "manipulation" (see also Shiffrin 2000 and, on nudges specifically, Anderson 2010). While I agree with Hausman and Welch (2010) on the conceptual point, some interventions that fall into this category are more than merely informational. For instance, *graphic* warning signs (such as the physical disfigurements shown on cigarette labels) are likely to have an independent psychological effect compared to non-graphic signs with the same content—an effect on implicit gut reactions over and above any that the information itself has on one's explicit beliefs. To the extent that they have these further implicit effects, some warnings and informational campaigns may still count as nudges by Hausman and Welch's own lights.

In sum, there are worries about whether certain nudges in fact fall under the auspices of libertarian paternalism, as well as "worries" about whether libertarian paternalism in general is actually paternalistic. However, the lion's share of criticism leveled at libertarian paternalism has alleged that the theory is not genuinely libertarian, and that some nudges amount to objectionable infringements on citizens' liberty, or autonomy.

## 6.3 Do Nudges Undermine Autonomy?

Public discussion of libertarian paternalism has been highly politicized, with conservative warnings that nudges amount to "brainwashing" and charges that David Cameron's Behavioral Insights Team is big brother's new "Ministry of Mind Control" (Townsend 2013). This is surely nudge paranoia, but much of the academic reception has been comparably critical.[3]

Jeremy Waldron (2014) notes that policymakers are subject to implicit biases just as much as the citizens they would attempt to nudge, and he warns against the intentional abuse of nudges and their potential to express disrespect for citizens. If policymakers truly respected citizens, they would make rational appeals designed to convince citizens to bring

[3]See, e.g., Bovens (2009), Gilles (2011), Rebonato (2012), and White (2013).

the same ends about of their own accord. Surely Waldron is right that some nudges have the potential to express attitudes of disrespect, superiority, condescension, and worse on the part of policymakers. But this possibility is far from unique to nudges. Any type of public policy has the potential to slight the dignity of the citizenry if motivated by the wrong intentions. Nudges might present more of a threat if so motivated than other types of intervention, but this would be due to some other, genuinely unique feature of nudges. And while policymakers surely are subject to implicit biases like everyone else, better to have those among the afflicted who are at least experts on the matter directing policy.

Til Grüne-Yanoff (2012) argues that nudges increase the government's "arbitrary power" over its citizens, which decreases their liberty (Pettit 1996). Officially, however, Thaler & Sunstein (2008: 3) claim that nudges are only used when "there is no such thing as a 'neutral' design." In such cases, no design involves *more* arbitrary power on the part of the government than any other (other things being equal), in which case Grüne-Yanoff's objection does not apply. The government is often simply stuck with a certain amount of arbitrary power that it must decide *which* way to exercise (including by omission). For instance, having some default is often unavoidable—you either have to have an "opt-in" or an "opt-out" organ donation system; you can't have neither. Often, nudging is unavoidable (Hausman and Welch 2010: 132, Shafir 2016). And when it is, at least one way of nudging is likely permissible, if not obligatory. Indeed, I think the point extends a step further: once we become aware of some instance of implicit interference with people's deliberation and behavior—that is, aware of the existence of some "inherent" threat to autonomy—we are forced to choose which way to nudge: either to cancel out the existing interference, or not.

Others worry not about nudges increasing the government's arbitrary power over citizens, but about nudges increasing policymakers' outright manipulativeness. This is why nudges have seemed to many to pose a threat to autonomy, specifically. Even if nudges preserve *liberty* in the sense of not taking away or making any options unduly costly, they do non-rationally "push" choosers one way rather than other ways or "shape" people's choices among options (Hausman and Welch 2010: 128; cf. Reiss 2013: 294 and Waldron 2014). In our terms, the real *prima facie* problem with nudges is that they are typically designed precisely to make people's implicit attitudes *hijack* their own practical reasoning, deliberation, and so choices. They're attempts to intervene on people's implicit attitudes in a way that leads people to change the way they deliberate and choose from what they otherwise would.

Hausman and Welch (2010: 130–1) come close to drawing just this conclusion:

> To the extent that [nudges] are attempts to undermine that individual's control over her own deliberation, as well as her ability to assess for herself her alternatives, they are *prima facie* as threatening to liberty, broadly understood, as is overt coercion. . . [T]here may be something more insidious about shaping choices than about open constraint. For example, suppose. . . subliminal messages were highly effective in influencing behavior. So the government might, for example, be able to increase the frequency with which people brush their teeth by requiring that the message, "Brush your teeth!" be flashed briefly during prime-time televi-

> sion programs. Influencing behavior in this way may be a greater threat to liberty, broadly conceived, than punishing drivers who do not wear seat belts, because it threatens people's control over their own evaluations and deliberation. . . [T]o the extent that it lessens the control agents have over their own evaluations, shaping people's choices for their own benefit seems to us to be alarmingly intrusive.

That is, in our terms, nudges not only influence behavior (as do beating and bypassing), but they do so precisely by affecting one's very reasoning and deliberation itself, and for this reason nudges seem particularly threatening to self-governance. Having less control over one's behavior is one thing; having less control over one's own explicit thoughts another. Ultimately, Hausman and Welch (2010: 134) think the benefits of nudging nonetheless outweigh the costs in terms of autonomy in many cases—the traditional justification for any kind of paternalism (even those that do interfere with citizens' option sets, like seatbelt laws).

Hausman and Welch and others, though, seem to assume that coercion and manipulation are problematic because they involve subjection to another agent's will, specifically. In place of citizens' judgments about their own good, they substitute policymakers' judgment about what's good for citizens instead (Hausman and Welch 2010: 129–130). This is why many think that nudges need not meet the same justificational bar when they'd essentially be replacing other actors' manipulation of citizens. Julian Reiss (2013: 298), for instance, claims that "[w]hen choice architecture is unavoidable we do not face the choice between the government interfering or abstaining from it, but rather the choice between government interfering or someone else." Marketers and advertisers, for instance, are increasingly looking to the findings of behavioral economicsto "push" customers into buying their products.[4]

In many cases, the relevant alternative to governmental nudging is intentional nudging by other third parties, like advertisers and the corporations they work for. As I argued in Chapter 2, though, what's actually autonomy-mitigating about hijacking and manipulation isn't being subject to another *agent*'s will, as such. Instead, intentional manipulation is just one salient way that actions and reasoning can be made less counterfactually dependent on agents' own values to a large extent. Other agents' intentional actions exhibit high degrees of counterfactual dependence, and so can have this type of effect. But other, completely "natural," non-agential sources—like one's own implicit attitudes—can exercise (intermediate degrees of) such influence, as well. Eldar Shafir (2016: 254) also notes that framing and contextual features can be just as "manipulative" in these ways. And in Chapter 5, we saw how unintentional social and geographic influences can "tap in" to people's implicit attitudes to produce effects exhibiting high counterfactual dependence. Perhaps none of these decrease the counterfactual dependence of citizens' actions on their values to quite the same degree as intentional manipulation by other agents, but the difference is one of degree, not kind. All of these "natural" forms of implicit, System 1 interference on System 2—what we might call *inherent nudges*—can be autonomy-mitigating in their own right to varying degrees.

[4]Sigmund Freud's nephew Edward Bernays was essential in developing modern advertising (that appeals to people's self-image rather than making rational appeals), in large part using ideas he first deployed while working in the U.S. propaganda department (Committee on Public Information) during WWI.

Thus, we're not only "forced to nudge" in cases where *someone* else will if the government does not, but also cases where something (*anything*) else will. Once we're aware of an inherent nudge, we are forced to choose whether to nudge another way (to remove or counteract the inherent nudge) or to do nothing. Either the government nudges, or it chooses to leave the inherent nudge in place. Even if it chooses not to choose, this is still a choice—namely, the choice to allow the inherent nudge to remain. Just as we are often "forced to choose" (Sartre 1956), so we are forced to nudge. Thus, blanket arguments that libertarian paternalism undermines autonomy do not succeed. The active interventions it advocates undermine autonomy *less* than any alternatives in many cases. The question is not whether nudges will be used and whether people's practical deliberation will be interfered with; deliberation is already routinely interfered with by people's own implicit attitudes in all the ways broached in previous chapters. Rather, the question in most cases is which nudge is *better*.

Fundamentally, the real issue over nudges is not about whether or not to interfere with autonomy—there will be interference either way—but only about the source and ultimate extent of the interference. In some cases, the choice is between governmental hijacking and hijacking by one's own System 1 (at least when there are no feasible alternative public policies that would simply appeal to one's System 2 through rational persuasion).[5] The question ultimately comes down to which nudge—the inherent nudge or the governmental intervention (and perhaps which governmental intervention)—undermines counterfactual dependence on one's values and autonomy *least*.[6] Many times, the governmental nudge is surely the lesser evil, undermining autonomy less than the pre-existing, inherent nudge. In these cases, there is no objection to nudging based on autonomy—autonomy is undermined *more* by the relevant alternative(s). In other cases, available interventions may come at a comparative cost to autonomy—but here, too, we need to know the extent of this difference in order to assess whether the tradeoff is worth whatever gains nudging would produce.

Thus, there is no cause for nudge paranoia, but we should also resist nudge hubris. To answer the normative question of whether any given nudge should be used or not, we must first answer the descriptive question of what effects it will have on the counterfactual dependence of citizens' choices on their values, and how this compares to the effect of alternative nudges on such counterfactual dependence.[7] Even if we don't always need any explicit calculation of these values, we'll generally need a decent appreciation of how they compare. However, as we've seen, these descriptive questions are often extremely complex, which should make us considerably humble in our normative aspirations. We may be forced to nudge, but doing so responsibly requires doing serious empirical background work on multiple levels.

[5] "[O]ne should distinguish between cases in which shaping increases the extent to which a person's decision-making is distorted by flaws in deliberation, and cases in which decision-making would be at least as distorted without any intentionally designed choice architecture" (Hausman and Welch 2010: 133).

[6] This is not to claim the counterfactual dependence framework alone provides a complete answer to these questions. At a minimum, people's actions and choices should also be non-deviantly dependent on—and so *express*—their own values. Many proposed nudges are designed to allow just this, removing obstacles between one's values and actions rather than substituting in others' values.

[7] On the normative question, see Sunstein (2015).

Behavioral game theorists, for instance, have shown how putting people into economic mindsets, period, can undermine moral ("pro-social") motivations. For example, imposing a fine on parents for dropping children off late at day care centers actually increases the number of late drop-offs (Gneezy & Rustichini 2000a, 2000b). Parents are at least somewhat disinclined to inconvenience day care workers when doing so seems rude or disrespectful. But when the choice is framed as an economic transaction, significantly more parents are willing to pay a fine in order for the opportunity to show up late. This case involves explicit incentives, but similar effects might be achieved by putting people into implicit economic mindsets—e.g., Vohs (2006) shows that exposure to dollar bill screen savers makes people behave more selfishly. Again, this is no general indictment of nudges, but it does suggest that even the seemingly least controversial financial nudges—aimed at getting more people to enroll in 401(k) plans, and the like—may be more complicated than initially meets the eye. Public policies can go wrong without proper research into what effects they'll have "on the ground" (and how they compare to existing, inherent effects). And if anything, the previous chapters show that the causal webs that nudges intervene on are often quite complex.[8]

## 6.4 Nudging Implicit Biases

The question of "debiasing" is nearly as old as the dual-processing literature itself, if not by the name "nudging" or discussed at the level of public policy (Fischhoff 1982). Among the least controversial nudges are perhaps those that correct for obvious financial heuristics and biases (like savings programs) and those meant to correct for racist, sexist, and other biases that nearly all people can agree everyone would be better off without. In this section, I discuss nudging the types of implicit racial bias that I argued in Chapter 5 are crucial to contemporary segregation and keeping poor urban blacks in poverty, before moving on to more controversial cases in the next: nudging the implicit attitudes of the poor, specifically.

Research on how to correct racist and sexist implicit associations, especially, has already begun in earnest. Interventions that have been found to reduce implicit bias include *approach-training*—coming to associate approach behaviors with the target, stigmatized individuals (Kawakami et al. 2005, 2007, Phills et al. 2011) or perhaps even mere social contact with and exposure to members of the group (Lowery, Hardin, and Sinclair 2001), *negation or denial training*—"saying no" to representations and expressions of the bias (Kawakami et al. 2000), and *exposure to counter-stereotypic exemplars* of the target group (Blair 2002)—e.g., exposure to admired black men (Dasgupta & Greenwald 2001, Gawronski et al. 2008).[9] A number of studies have also investigated the use of implementation intentions—intentions

[8]For instance, Titmuss (1970) predicts that monetary incentives will also decrease blood donations, but empirical work suggests that they actually have the opposite effect (Mellström & Johannesson 2008, Lacetera & Macis 2010, Lacetera, Macis, and Slonin 2012). Of course, good research is also no guarantee that policies won't lead to unintended effects, but such background does decrease the likelihood of unexpected effects.

[9]Recall from Ch. 5 that admirable black male role models are scarce for precisely those who need them most—poor young black men. For the long-term effects of these types of training, see Devine et al. (2012).

with a conditional form ("I intend to $A$ if circumstances $C$ arise") or to perform a particular action in the presence of specific triggers. And there is evidence that forming implementation intentions not to act on one's implicit biases is effective (Gollwitzer et al. 2005, Mendoza et al. 2010, Moskowitz et al. 1999, 2011, Stewart & Payne 2008, and Webb et al. 2010).

There's also been extensive talk about how the law might intervene to reduce implicit bias and its effects using the dual-processing research. For instance, Dasgupta & Greenwald (2001) show that IAT biases in favor of whites are diminished after exposure to photographs of admired black men (e.g., Martin Luther King). And Jolls & Sunstein (2006b) discuss how negative and positive imagery might be legally regulated in workplaces. For example, stereotypical or negative images of minority groups might be prohibited, and incentives might be created to display positive images of minority individuals as defenses against employers' vicarious liability for Title VII violations (cf. Jolls & Sunstein 2006a).

Advisable as these measures may be, they invite another cautionary note. Images of admired individuals from a given minority group might decrease implicit biases against that group, but such images also portray these individuals' *eyes*, and as noted above, the presence of eyes in one's environment has other effects on behavior. In this case, the additional, unintended effect (of eyes) may be socially beneficial, but this won't always be the case. In general, policymakers should be careful to consider *all* potential implicit effects of proposed nudges (at least those on which a proposed nudge and the inherent, existing influences differ). Otherwise, a nudge's unintended negative side-effects could outweigh its intended benefits. Again, the issue is not whether to nudge or not, and in some cases we may be confident that it's better to intervene in some way rather than allowing inherent nudges to remain in place. But even in these cases, we must still determine *which* intervention to use—e.g., whether to nudge using pictures of admired minority members or negation training.[10]

More ambitiously, Kang & Banaji (2006) argue that implicit bias may ground a novel argument for affirmative action (see also Jolls & Sunstein 2006b). If the implicit biases of college admissions personnel typically lead them to choose white candidates more often than equally qualified black candidates (e.g., Dovidio & Gaertner 2000 and Hodson, Dovidio, and Gaertner 2002), then giving explicit institutional preference to black applicants may counteract this effect. Again, this may be the correct normative conclusion. But this is another example of where public policy at least needs to be informed by a broader body of research than social psychology on implicit attitudes. In this case, there's some sociological work suggesting that affirmative action may have exactly the opposite long-term consequences as those intended. William Julius Wilson (1987/2012), at least, has argued that affirmative action only tends to be accessible to—and so only tends to help—middle- and upper-class blacks. Hence, it contributes to the exodus of middle- and upper-class blacks from the inner city, leaving behind higher concentrations of (even) poorer blacks, thereby "ratcheting up" the impact of poverty (Chapter 5). This is not to agree with Wilson. But again, policymak-

[10]Many researchers have warned that trying to explicitly reason people out of their implicit biases may have "rebound effects," actually strengthening the biases (Follenfant & Ric 2010), similar to trying not to think about white bears (Wegner 1989). If associations are blind to negation (Ch. 2), this is especially likely.

ers need to grapple with these sorts of ecological considerations, beyond those of laboratory psychology and behavioral economics, in order to make responsible policy decisions.

Psychologists and sociologists should also attend to what might be done to nudge realtors, landlords, and public housing officials out of their implicit biases, which I argued in the last chapter are crucial factors in keeping the ghetto poor in poverty. This is especially true since the only current means of legal rectification for housing discrimination is to prove beyond a reasonable doubt in a court of law that one has been discriminated against—an extremely difficult task when such biases are often only apparent in the statistical aggregate.

## 6.5 Nudging the Implicit Attitudes of the Poor

Some philosophical work has begun connecting the psychological research on implicit attitudes to social theory, especially theories of racism (e.g., Valian 2005, Kelly & Roedder 2008, Faucher & Machery 2009, Anderson 2010, Machery et al. 2010, and Madva 2016). Kelly & Roedder (2008), for instance, suggest that we should change implicit attitudes by explicitly engineering people's environments. Most of this work has not, however, focused on nudging the implicit attitudes of the poor themselves, or connected research on nudges and implicit attitudes to the sociological work on poverty that was my focus in Chapter 5.

Some nudges have targeted the implicit attitudes of the urban poor by intervening on inner-city environments. David Cameron's Behavioral Insights Team, for instance, proposed to change the physical environment of the Hillington Square housing projects in King's Lynn, Norfolk in ways that would reduce crime and improve residents' sense of community. Specifically, they recommended the removal of "dead spaces" used by drug dealers and the homeless, as well as the removal of dead ends from walkways and "dark corners" more generally. These proposals seem relatively innocuous, but the fact that interventions on geographic and social factors have the potential to affect not only citizens' behavior and deliberation, but also what they value and their very identities itself, raises special concerns.

Many nudges only affect one's mere "first-order" associations, those not involving the concept SELF. These raise fewer worries about autonomy. However, some nudges have the potential to affect one's implicit self-associations (ISAs), and so what one identifies with—one's "practical identity" or the "description" under which one values oneself on the implicit level (Chapter 4). Concerns about autonomy surrounding nudges come into sharpest focus in these cases: when we consider nudging citizens' very values and implicit self-worth (whether the poor or others). As suggested in Chapters 4 and 5, one's values and self-worth are intimately tied to one's culture, and culture is in many ways realized by ISAs.

Thaler and Sunstein (2008: 53) recognize the connection between implicit associations and culture: one of the most effective ways to nudge is through interpersonal influence, and small cultural nudges can lead to large social changes.[11] Many times, members of one's society nudge one unintentionally, leading to *implicit conformity* (Thaler and Sunstein 2008: 64). Solomon Asch's (1955) studies are the classic, in which participants' judgments about

[11]See also Ross & Nisbett (1991), Cialdini (2000), and Sunstein (2003).

the (comparative) lengths of several lines drawn on a chalkboard are significantly influenced by other participants' (and confederates') judgments. Sherif (1937) finds that groups tend to converge on stable, shared judgments, but that these are highly subject to initial arbitrary influences.[12] Similarly, Salganik, Dodds, and Watts (2006) find that song downloads from an artificial music website (an "artificial cultural market") are heavily influenced by the number of previous downloads (social popularity), though the ultimate popularity of each song is highly sensitive to initial arbitrary conditions (cf. Jacobs & Campbell 1991). Such effects of "social contagion," or of implicitly "following the herd" extend to teen pregnancy (Akerlof, Yellen, and Katz 1996), obesity (Christakis & Fowler 2007), the academic effort of college students—whose behavior varies depending on which dorm and even roommates they have (Sacerdote 2001), and even the voting patterns of federal judges (Sunstein et al. 2006). Shiller (2000, 2008) argues that social conformity is the main cause of economic over-speculation, including that which caused the financial collapse of 2008.

Crutchfield (1955) shows that there are also conformity effects on professed values. When asked whether economic recession, educational facilities, mental health, crime and corruption, or subversive activities are the most important problem facing the country, 19% of individuals chose subversive activities when they answered the question privately, but 48% said this was the important problem when the rest of their group did. Only 19% of participants in private agreed that society had a right to suspend free speech when it was perceived as threatening, but 58% when the rest of their group said it was permissible. Presumably, no participant explicitly reasons to themselves that their answer should be different because other people say so; instead, others' judgments seem to hijack one's own reasoning via implicit System 1 pressures toward conformity. Once again, one's conception of *THE GOOD* is framed—hijacked out of the control of one's actual values. Government might try to correct for such conformity effects, perhaps in part by leveraging the same mechanisms.

Some nudges already attempt to shape implicit values using these social conformity mechanisms, as Thaler and Sunstein suggest. Consider the wildly successful "Don't Mess With Texas" anti-littering campaign. Well aware that their target audience was 18-24 year old males, Texas officials enlisted Dallas Cowboys football players and country music stars who "smashed beer cans in their bare hands, and growled 'Don't mess with Texas!'" in campaign ads—ads meant to contain "a tough-talking slogan that would also address the unique spirit of Texas pride" (Thaler and Sunstein 2008: 60). In other words, Texas state officials attempted to get citizens to stop littering by implicitly associating doing so with "Texas pride" and local culture, which Texans from the relevant demographic in turn had strong implicit self-associations with.[13] Consistent with the suggestion in Chapter 4 that ISAs tend to have an especially strong impact on behavior (since they involve coming to

[12]Recall the conclusions of Ch. 3 on "arbitrary coherence." Often, the "conformity experiments" are interpreted as showing that people explicitly *reason* that because others' judgments are different, their own perceptual judgments must be mistaken, leading them to give adjusted answers. Perhaps it's true that some participants explicitly infer that they can't trust their own eyes, but more plausibly, a large share of the effect is driven by implicit thinking of the sorts discussed in Chapters 2 and 3. See Murray (2015, forthcoming).

[13]Recall the "enemy-of-my-enemy-is-my-friend" effect from Ch. 4.

see one's own implicit self-worth as contingent on the worth of the thing in question), the campaign reduced litter by 29% in its first year and by 72% within 6 years.[14]

Employing nudges that add to one's existing set of ISAs in these ways is surely sometimes permissible, if not obligatory. But measures that would modify or remove extant ISAs from citizens' cognitive economies are likely more controversial. I close with a particularly striking case that dovetails with the last chapter and our main moral ecological example.

As discussed in Chapter 5, many sociologists have claimed that a culture of poverty or subculture of scarcity—in parts a type of honor culture—exists in many American inner cities. More controversially, aspects of this culture are part of what perpetuate cycles of poverty in these neighborhoods. And many have hoped that certain public policies, such as housing voucher programs, would stymie these phenomena. However, many recipients of Section 8 and other vouchers that subsidize housing only use their vouchers to relocate to other socially disadvantaged neighborhoods (Sampson 2012, Venkatesh 2000). This effect is likely due in part to neighborhood (type) self-associations; ghetto residents aren't comfortable moving to non-disadvantaged neighborhoods—they feel as though they wouldn't know how to live there and perhaps can't really imagine what it would be like to do so. Instead, they feel as though they're only comfortable or belong in comparatively poor and disadvantaged neighborhoods because they associate themselves (have ISAs) with such neighborhoods.[15] Darrah & DeLuca's (2014) report on how the *Baltimore Mobility Program (BMP)* fared differently from many other voucher programs supports this interpretation.

Unlike other voucher programs like the *Moving to Opportunity (MTO)* experiment, the BMP changed the housing *default*, requiring participating families to move to low-poverty, non-racially-segregated neighborhoods for (at least) the first year of the program (after which families could move anywhere they chose. Thus, the BMP isn't a nudge proper—at least for the first year—since it does restrict participants' choice sets.) The BMP also differed from other voucher programs by giving families extensive counseling, which, as Darrah & DeLuca describe it, mainly reinforced the effects of having to live outside concentrated disadvantage for a year—raising parents' "expectations" about what one's neighborhood, home, and schools could possibly provide by shifting parents' "residential choice frameworks." In particular, Darrah & DeLuca (2014: 356) claim that exposure to a new neighborhood (type) and counseling generated "awareness of the benefits to specific attributes" of different neighborhood options, in effect adding new neighborhoods to participants' choice sets.

On this interpretation, many ghetto residents simply do not see less-disadvantaged neighborhoods as genuine options without sufficient counseling of the kind provided by BMP. Krysan & Bader (2009) claim that many of the severely disadvantaged have "neighborhood blind spots"—i.e., they simply never consider certain neighborhoods as places that they might potentially move to. As one of Darrah & DeLuca's (2014: 364) participants puts it: "living in the ghetto... if this is all you see, this is all you know." Results indicate that 60%

[14]Once again, while the program was quite successful, there's room for worry about unintended consequences: namely, strengthening other associations with "Texas pride" and culture, like machismo.

[15]See also Logan, Stults, and Farley (2004), Quillian (2012), and Sampson & Sharkey (2008).

of sample BMP movers experienced changes in their choice frameworks after the first year of the program, in large part because it served to erase neighborhood blind spots. Nudges that simply overcome and fill in these types of blind spots should be particularly unobjectionable. Nudges that overcome the "scarcity mindset" and narrowing of attention (responsible for overborrowing) induced by poverty more generally, documented by Mullainathan and Shafir (2013) and discussed in the last chapter, should also be uncontroversial.

However, there's another (not mutually exclusive) interpretation that might be given of these particular BMP results, as well. Darrah & DeLuca (2014: 374–6) themselves suggest that BMP increased the number of neighborhoods participants felt "familiar" with. Counseling and other support helped "participants to imagine themselves living—and thriving—in unfamiliar places" (Darrah & DeLuca 2014: 362). This suggests that BMP may have changed participants' implicit self-associations—creating new ISAs with affluent neighborhoods and decreasing participants' implicit identification with ghetto neighborhoods.

Darrah & DeLuca (2014: 376) conclude that "additional supports and resources, such as counseling and changes to *choice architecture* may be necessary" to actually make housing voucher programs successful in relieving concentrated disadvantage. This kind of invocation of Thaler and Sunstein (2008) is becoming commonplace in housing policy discussions. Hall, Galvez, and Sederbaum (2014) suggest that "low-income consumers may need 'nudges' or other situational interventions that will encourage them to follow through with their stated goals or preferences," and de Souza Briggs, Popkin, and Goering (2010) implicate *Nudge* directly in the lessons to be drawn from the MTO studies. De Souza Briggs (2008) suggests making "better choices the defaults or starting points," and de Souza Briggs, Comey, and Weismann (2010: 420) explicitly advocate that "[s]ome reforms could 'change the default,' in the language of choice architecture (Thaler and Sunstein 2008)... to use the voucher to considering recommended options first." In other words, policy experts themselves are starting to suggest that we might make moving to non-socially-disadvantaged neighborhoods the default or status quo for those who receive housing vouchers, such that recipients would have to "opt out" to remain in ghetto neighborhoods. In effect, these experts are recommending hijacking as a (partial) solution to the cycles of poverty endemic to U.S. ghettos.

Changing the default to non-disadvantaged neighborhoods without making them required would turn a given housing voucher program into a nudge. And it seems quite likely that such nudges would affect people's ISAs—which neighborhoods and so ways of life they identify or self-associate with. If sufficient numbers of residents were nudged in these ways, it might affect the culture of the inner city itself. We know that the effects of changing default options can be dramatic elsewhere—for instance, in organ donation. And once these effects become the norm—with most ghetto residents preferring to move to other neighborhoods, in this instance—implicit social conformity will only increase the default's influence. Further empirical (and conceptual) work on just what effects housing voucher defaults would have are certainly needed. But in the hypothetical limiting case, such changes to the default (perhaps in conjunction with other policies aimed at desegregation) might lead to the eventual dissolution of ghetto neighborhoods and their attendant culture entirely. Perhaps most voucher recipients would choose the default, and perhaps as more people came to identify with non-

disadvantaged neighborhoods, they would identify less with the "code of the street" and the subculture of scarcity. Eventually, cultural pressure might completely quash these values and behavioral patterns as viable ways of life. This, I believe, is what's genuinely worrisome about (some) nudges: their potential to have large changes on what agents implicitly value, care about, and "where agents (implicitly) stand" on fundamental practical matters.[16]

To some, the extinction of U.S. ghettos and their associated cultural characteristics might seem an auspicious prospect. Elizabeth Anderson (2010), for instance, argues that social justice requires residential (and other forms of) desegregation, and she includes affirmative action and housing voucher programs in her proposals for how to bring about such reform. The existence of overwhelmingly black ghettos is inconsistent with such proposals for desegregation. Others, of course, favor investment in ghetto neighborhoods themselves, rather than moving their residents elsewhere (e.g., Shelby 2014, 2016). Again, I do not mean to take any normative stand here. I do not suggest that nudges which change the defaults of housing voucher programs should not be employed (nor that they should), nor do I mean to endorse any form of cultural conservationism. But the example does highlight the importance of the threat to autonomy that nudges capable of affecting culture and ISAs raise.

As Appiah (2005: Ch. 5) and Shelby (2016: 91) note, governmental interventions that teach citizens new skills or cultural traits, or that correct mistaken factual beliefs, are one thing, but interventions which attempt to change people's values and very identities are quite another, and must meet a higher justificational bar. This is because one's cultural traits and values (including, we can add, those that are implicit) determine one's sense of self-worth, and so the possible sources of achieving such worth (Shelby 2016: 96-100). It would be inhumane to simply take away people's only viable source of such satisfaction, and this may well be the position that undermining the subculture of scarcity and code of the street would put many ghetto residents in. If nudged to leave the ghetto and cease self-associating with these values and behavioral patterns, former ghetto residents may be left without any means to attain implicit self-worth—to live a life worth living, by these citizens' own lights.

It might be argued that nudges which affect culture and implicit identity do not simply remove one culture and set of values, but that they replace it with another—in this case, "mainstream" values. If the (set of) options a nudge puts in place (including potential sources of self-worth) are equally valuable with the (set of) options replaced, it may not be objectionable that the nudge *changes* one's autonomy, so long as it does not *reduce* it.[17] In other words, it might be argued that changing one's practical identity to something different isn't objectionable, so long as that identity and its sources of self-worth are just as good (or better) as what they replace. Perhaps even if some cultural nudges interfere with what people self-associate with at $t_1$, they might promote autonomy by changing, and then being in accord with, what people self-associate with at $_2$. In that case, many cultural nudges that change one's ISAs would be less objectionable than they might otherwise seem.[18]

[16]These effects on implicit values might then percolate up to affect explicit values and practical identities.

[17]There may also be other, non-autonomy-based objections to such interventions (Shelby 2007, 2016).

[18]Doris (2015: Ch. 6), for instance, argues that in some cases *pre-hoc* rationalization of one's choices and

Ultimately, the question of whether "equal or better" alterations to what one identifies with are morally objectionable turns on whether there are historical conditions on autonomy, as discussed at the close of Chapter 3 (see Fischer & Ravizza 1998, Vargas 2006, and Watson 2001). If *ahistoricists* are right that autonomy is determined only by current, "time-slice" properties of an agent, then changing the nature of one's autonomy by changing one's implicit values may be permissible. If *historicists* are right that properties regarding how an agent came to have the set of values and identity she currently has are also relevant, then such changes may be impermissible. I won't try to settle the debate here, only note that it's another place where traditional philosophical work on autonomy can inform, and be informed by, important real-world issues of the sort broached in this and the last chapter. Indeed, to return to the main, overarching argument introduced in Chapter 1, I hope it's by now clear that we cannot understand how autonomy works "on the ground," historically or otherwise, without understanding the types of considerations raised by moral ecology.

## 6.6 Conclusion

Questions about when and how we should intervene in citizens' lives are ultimately questions for normative ethicists, political philosophers, public policy experts, and politicians. But in many cases, the question is not, as Hausman and Welch (2010) put it: "to nudge or not to nudge?" The question is instead: "which way to nudge?" Shafir (2016) is correct that we're often "manipulated," whether intentionally or not—that manipulation of this sort is an unavoidable part of life—and that we need to make the most of it. We need to "manipulate" in the best ways, with the best information available—from all the relevant sciences.

Many nudges no doubt promote autonomy and other values—like equality, quality of life, and the like—better than their inherent (and other) alternatives, and so are (at least) permissible. This is especially true of nudges expressly designed to remove known sources of hijacking and other forms of implicit interference with autonomy—nudges that put agents' reasoning and behavior back in the hands of agents' values, and which do so better than any relevant alternatives. *Often, the best way to correct for the hijacking of reason is further (better informed) hijacking.* Nudges that carefully employ such hijacking should not be considered sources of causal deviance, but rather as removing obstacles in the causal chain between one's choices and actions and what one values and cares about. Still, special caution is called for in cases where nudges might affect one's values themselves. Even the most liberal governments unavoidably engage in this kind of "soul-making" (Appiah 2005, Shelby 2016), public education being perhaps the primary and in many ways most controversial example. This is not a reason for the government to play a more restricted role in the lives of its citizens, though, but for it to avail itself of all the expertise available—not only from cognitive scientists and behavioral economists, but also from sociologists and philosophers—ideally, from those who incorporate the insights of all these fields—that is, from moral ecologists.

beliefs of the sort we've focused on—e.g., positive illusions of control—may actually facilitate agency.

Shafir (2016: 257) goes too far in suggesting that we need not worry about the potential for nudges to threaten autonomy because autonomy is largely illusory. The "inherent nudges" of advertisers, the ambient environment, and our own and others' unconscious do not undermine autonomy entirely, but only to varying degrees. What we need to determine is instead the comparative extent to which nudges interfere with and bolster autonomy, as well as how to balance out gains and losses in autonomy with gains and losses with respect to other values. The counterfactual dependence framework provided by interventionist theories of causation (Chapter 2) provides a way of understanding what we've seen are often highly complex, interconnected effects on autonomy. Determining just what values the variables and connection weights in these models take is, in many cases, highly difficult, but we can make approximations that are often "good enough for government work."

To the extent that moral thought and behavior is socially and geographically-situated, a certain lack of specificity in the models needed to make responsible public policy decisions is unavoidable. We should only seek the precision that a given subject matter allows for (Aristotle 2004). And we must treat policy decisions as partly experimental—open to revision and not closed by (bipartisan) political prejudice when it becomes apparent that more beneficial programs might be implemented. In this dissertation, I've attempted to trace out some of the most rudimentary connections that should prove useful for the field of moral ecology—the study of moral interactions between agents within social and geographic environments. Even if the details of my own proposals require revision, the importance of the connections between these different ideas and fields of study stands. Philosophers working in normative ethics and political philosophy have a role to play in "government work"—in ethical and political questions that have a real impact on the lives of citizens and nations—and those who aspire to such work should be able to avail themselves of a mature moral ecology.

# References

Akerlof, G., Yellen, J., and Katz, M. 1996. "An analysis of out-of-wedlock childbearing in the United States." *Quarterly Journal of Economics*, 111: 277–317.

Akerlof, G., & Shiller, R. 2015. *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton, NJ: Princeton University Press.

Allport, G. 1954. *The Nature of Prejudice*. Reading, MA: Addison-Wesley.

Anderson, Elijah. 1990. *Streetwise: Race, Class, and Change in an Urban Community*. Chicago: University of Chicago Press.

Anderson, Elijah. 1999. *Code of the Street: Decency, Violence, and the Moral Life of the Inner City*. New York: Norton.

Anderson, Elizabeth. 2010. *The Imperative of Integration*. Princeton: Princeton University Press.

Appiah, K. 2005. *The Ethics of Identity*. Princeton: Princeton University Press.

Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., and Amadori, A. 2008. "Predicting the Vote: Implicit Attitudes as Predictors of the Future Behavior of Decided and Undecided Voters." *Political Psychology*, 29: 369–387.

Ariely, D., & Wallsten, T. 1995. "Seeking subjective dominance in multidimensional space: An explanation of the asymmetric dominance effect." *Organizational Behavior and Human Decision Processes*, 63: 223–232.

Ariely, D., & Carmon, Z. 2003. "Summary assessment of experiences: the whole is different from the sum of its parts." In G. Loewenstein, D. Read, and R. Baumeister (Eds.), *Time and decision: Economic and psychological perspectives on intertemporal choice*, (pp. 323–350). New York: Russell Sage.

Ariely, D., Loewenstein, G., and Prelec, D. 2003. "Coherent Arbitrariness: Stable Demand Curves Without Stable Preferences." *Quarterly Journal of Economics*, 118: 73–105.

Ariely, D., Loewenstein, G., and Prelec, D. 2006. "Tom Sawyer and the Construction of Value." *Journal of Economic Behavior and Organization*, 60: 1–10.

Ariely, D. 2008. *Predictably Irrational: The Hidden Forces that Shape Our Decisions*. New York: Harper.

Ariely, D. and Norton, M. I. 2008. "How Actions Create – Not Just Reveal – Preferences." *Trends in*

*Cognitive Sciences*, 12: 13–16.

Aristotle. Thomson, J. (Ed.). 2004. *The Nicomachean Ethics*. New York: Penguin.

Arpaly, N. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.

Arpaly, N. 2005. "How It Is Not "Just Like Diabetes": Mental Disorders and the Moral Psychologist." *Philosophical Issues*, 15: 282–298.

Arpaly, N. 2006. *Merit, Meaning, and Human Bondage*. Princeton University Press.

Arpaly, N., & Schroeder, T. 1999. "Praise, Blame, and the Whole Self." *Philosophical Studies*, 93: 161–188.

Arpaly, N., & Schroeder, T. 2012. "Deliberation and Acting for Reasons." *Philosophical Review*, 121: 209–39.

Arplay, N., & Schroeder, T. 2014. *In Praise of Desire*. Oxford: Oxford University Press.

Arrow, K. 1950. "A difficulty in the concept of social welfare." *Journal of Political Economy*, 58: 328–346.

Arrow, K. 1951. *Social choice and individual values*. New York: John Wiley & Sons.

Asch, S. 1955. "Opinions and social pressure." *Scientific American*, 193: 31–35.

Banaji, M., and Greenwald, A. 2013. *Blindspot*. New York: Delacorte Press.

Bargh, J., Chen, M., and Burrows, L. 1996. "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action." *Journal of Personality and Social Psychology*, 71: 230–44.

Bargh, J. 2008. "Free will is un-natural." In J. Baer, J. Kaufmann & R. Baumeister (Eds.), *Are we free? Psychology and free will*, (pp. 128-154). New York: Oxford University Press.

Baron, J. 2007. *Thinking and deciding* (4th Ed.). Cambridge: Cambridge University Press.

Bateson, M., Nettle, D., and Roberts, G. 2006. "Cues of being watched enhance cooperation in a real-world setting." *Biology Letters*, 2: 412–414.

Baumeister, R., Bratslavsky, E., Muraven, M., and Tice, D. 1998. "Ego depletion: Is the active self a limited resource?" *Journal of Personality and Social Psychology*, 74: 1252–1265.

Baumeister, R., & Masicampo, E. 2010. "Conscious thought is for facilitating social and cultural interactions: How mental simulations serve the animal-culture interface." *Psychological Review*, 117: 945–971.

Bechara, A. 2005. "Decision making, impulse control and loss of willpower to resist drugs: A neurocognitive perspective." *Nature Neuroscience*, 8: 1458–1463.

Becker, G. 1968. "Crime and Punishment." *Journal of Political Economy*, 76: 196–217.

Beggan, J. 1992. "On the social nature of nonsocial perception: The mere ownership effect." *Journal of Personality and Social Psychology*, 62: 229–237.

Bermúdez, J. 2009. *Decision Theory and Rationality*. Oxford: Oxford University Press.

Bertrand, M. and Mullainathan, S. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94: 991–1013.

Bessenoff, G., and Sherman, J. 2000. "Automatic and controlled components of prejudice toward fat people: Evaluation versus stereotype activation." *Social Cognition*, 18: 329–53.

Blair, I. 2002. "The malleability of automatic stereotypes and prejudice." *Personality and Social Psychology Review*, 3: 242–261.

Blair, I., Judd, C., and Chapleau, K. 2004. "The influence of Afrocentric facial features in criminal sentencing." *Psychological Science*, 15: 674–679.

Bratman, M. 2007. *Structures of Agency*. Oxford: Oxford University Press.

Bratman, M. 2009a. "Intention, Belief, Practical, Theoretical." In S. Robertson (Ed.), *Spheres of Reason: New Essays on the Philosophy of Normativity*, (pp. 29–61). Oxford: Oxford University Press.

Bratman, M. 2009b. "Intention, Practical Rationality, and Self-Governance." *Ethics*, 119: 411–443.

Bratman, M. 2014. *Shared Agency*. Oxford: Oxford University Press.

Brehm, J. 1956. "Postdecision changes in the desirability of alternatives." *Journal of Abnormal and Social Psychology*, 52: 384–389.

Brendl, C., Chattopadhyay, A., Pelham, B., and Carvallo, M. 2005. "Name letter branding: Valence transfers when product specific needs are active." *Journal of Consumer Research*, 32: 405–415.

Brewer, M., and Gardner, W. 1996. "Who is this 'we'? Levels of collective identity and self representations." *Journal of Personality and Social Psychology*, 71: 83–93.

Brigham, J. 1993. "College students' racial attitudes." *Journal of Applied Social Psychology*, 23: 1933–1967.

Brownstein, A. 2003. "Biased Predecision Processing." *Psychological Bulletin*, 129: 545–568.

Brownstein, M., and Saul, J. (Eds.). 2016. *Implicit Bias & Philosophy, Vol. 2: Moral Responsibility, Structural Injustice, and Ethics*. Oxford: Oxford University Press.

Bruner, J., and Goodman, C. 1947. "Value and need as organizing factors in perception." *Journal of Abnormal and Social Psychology*, 42: 33–44.

Burnham, T., & Hare, B. 2007. "Engineering human cooperation – Does involuntary neural activation increase public goods cooperation?" *Human Nature*, 18: 88–108.

Cadinu, M., & Rothbart, M. 1996. "Self-anchoring and differentiation processes in the minimal group setting." *Journal of Personality and Social Psychology*, 70: 661–677.

Camerer, C., 1992. "The rationality of prices and volume in experimental markets." *Organizational Behavior and Human Decision Processes*, 51: 237–272.

Cameron, Z., & Ariely, D. 2000. "Focusing on the forgone: How value can appear so different to buyers and sellers." *Journal of Consumer Research*, 27: 360–370.

Campbell, J. 2007. "An Interventionist Approach to Causation in Psychology." In A. Gopnik and L. Schulz (Eds.), *Causal Learning: Psychology, Philosophy and Computation*, (pp. 58-66). Oxford: Oxford University Press.

Campbell, J. 2010. "Control Variables and Mental Causation," *Proceedings of the Aristotelian Society*, 110: 15–30.

Capes, J. 2013. "Mitigating Soft Compatibilism." *Philosophy and Phenomenological Research*, 87: 640–663.

Carlson, K., Meloy, M., Russo, J. 2006. "Leader-driven primacy: Using attribute order to affect consumer choice." *Journal of Consumer Research*, 32: 513–518.

Chaiken, S., and Trope, Y. (Eds.). 1999. *Dual-process theories in social psychology.* New York: Guilford Press.

Chaxel, A. Caroline; Russo, J. Edward; Kerimi, Neda. 2013. "Preference-driven biases in decision makers' information search and evaluation." *Judgment and Decision Making*, 8.5 : 561–576.

Chaxel, A. 2015. "How do stereotypes influence choice?" *Psychological Science*, 26: 641–645.

Christakis, N., & Fowler, J. 2007. "The spread of obesity in a large social network over 32 years." *New England Journal of Medicine*, 357: 370–379.

Cialdini, R. 2000. *Influence: Science and Practice*, 4th Ed. Needham Heights, MA: Allyn and Bacon.

Clark, K. 1965. *Dark Ghetto: Dilemmas of Social Power.* New York: Harper and Row.

Clark, K., and Clark, M. 1941/1958. "Racial Identification and Preference in Negro Children." In Maccoby, Newcomb, and Hartley (Eds.), *Readings in Social Psychology*, (pp. 602–611). New York: Holt Reinhard Co.

Coates, J., & Swenson, P. 2013. "Reasons-Responsiveness and Degrees of Responsibility." *Philosophical Studies*, 165: 629–645.

Coppin, G., Delplanque, S., Cayeux, I., Porcherot, C., and Sander, D. 2010. "I'm no longer torn after choice: How explicit choices implicitly shape preferences of odors." *Psychological Science*, 21: 489–493.

Coppin, G., Delplanque, S., Porcherot, C., Cayeux, I., and Sander, D. 2012. "When flexibility is stable: Implicit long-term shaping of olfactory preferences." *PLoS One*, 7: e37857.

Correll, J., Park, B., Judd, C., and Wittenbrink, B. 2002. "The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals." *Journal of Personality and Social Psychology*, 83: 1314–1329.

Cushman, F., Young, L., & Hauser, M. 2006. "The Role of Conscious Reasoning and Intuitions in Moral Judgment: Testing Three Principles of Harm." *Psychological Science*, 17: 1082–9.

Darrah, J., & DeLuca, S. 2014. "'Living here has changed my whole perspective': How escaping inner-city poverty shapes neighborhood and housing choice." *Journal of Policy Analysis and Management*, 33: 350–384.

Dasgupta, N., & Greenwald, A. 2001. "On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals." *Journal of Personality and Social Psychology*, 81: 800–814.

Davidson, D., McKinsey, J., and Suppes, P. 1955. "Outlines of a Formal Theory of Value." *Philosophy of Science*, 22: 140–160.

Davidson, D., & Suppes, P. 1957. *Decision-Making: An Experimental Approach.* Stanford: Stanford University Press.

Davidson, D., & Marschak, J. 1959. "Experimental tests of a stochastic decision theory." In C. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories*, (pp. 233–269). New York: Wiley.

Davidson, D. 1963. "Actions, Reasons, and Causes." *Journal of Philosophy*, 60: 685–700. Reprinted in *Essays on Actions & Events*: 1980, Oxford: Oxford University Press.

Davidson, D. 1974. "Psychology as Philosophy," in S. C. Brown (Ed.), *Philosophy of Psychology*, 1974, London: Macmillan. Reprinted in *Essays on Actions & Events*: 1980, Oxford: Oxford University Press.

de Souza Briggs, X. 2008. "Strengthening Housing Opportunity through the Housing Choice Voucher Program." *Testimony before the National Commission on Fair Housing and Equal Opportunity*, Boston, Massachusetts: September 22.

de Souza Briggs, X., Comey, J., and Weismann, G. 2010. "Struggling to stay out of high-poverty neighborhoods: housing choice and locations in Moving to Opportunity's first decade." *Housing Policy Debate*, 20: 383–427.

de Souza Briggs, X., Popkin, S., Goering, J. 2010. *Moving to Opportunity: The Story of an American Experiment to Fight Ghetto Poverty.* Oxford: Oxford University Press.

Deery, O., & Nahmias, E. 2017. "Defeating Manipulation Arguments: Interventionist Causation and Compatibilist Sourcehood." *Philosophical Studies*, 174: 1255–1276.

Desmond, M. 2016. *Evicted: Poverty and Profit in the American City.* New York: Crown Publishers.

Deutsch, R., Gawronski, B., and Strack, F. 2006. "At the Boundaries of Automaticity: Negation as Reflective Operation." *Journal of Personality and Social Psychology*, 91: 385–405.

Devine, P., Plant, E., Amodio, D., Harmon-Jones, E., and Vance, S. 2002. "The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice." *Journal of Personality and Social Psychology*, 82: 835–848.

Devine, P., Forscher, P., Austin, A., and Cox, W. 2012. "Long-term reduction in implicit race bias: A prejudice habit-breaking intervention." *Journal of Experimental Social Psychology*, 48: 1267–1268.

Dewey, J. 1957. *Outlines of a Critical Theory of Ethics.* New York: Hillary House.

Dietrich, F., & List, C. 2013a. "A Reason-Based Theory of Rational Choice." *Noûs*, 47: 104–134.

Dietrich, F., & List, C. 2013b. "Where do preferences come from?" *International Journal of Game Theory*, 42: 613–637.

Dijksterhuis & Aarts, H. 2010. "Goals, attention, and (un)consciousness." *Annual Review of Psychology*, 61: 467–490.

Ditto, P., Pizarro, D., and Tannenbaum, D. 2009. "Motivated moral reasoning." In B. Ross (Series Ed.) and D. Bartels, C. Bauman, L. Skitka, and D. Medin (Eds.), *Psychology of learning and motivation, Vol. 50: Moral judgment and decision making*, (pp. 307–338). San Diego, CA: Academic Press.

Ditto, P. & Liu, B. 2011. "Deontological dissonance and the consequentialist crutch." In M. Mikulincer and P. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil*, (pp. 51–70). Washington, D.C.: American Psychological Association.

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

Doris, J. and The Moral Psychology Research Group (Eds.). 2010. *The Moral Psychology Handbook.* Oxford: Oxford University Press.

Doris, J. 2015. *Talking to Our Selves: Reflection, Skepticism, and Agency.* Oxford: Oxford University Press.

Dovidio, J., Kawakami, K., Johnson, C., Johnson, B., and Howard, A. 1997. "On the nature of prejudice: Automatic and controlled processes." *Journal of Experimental Social Psychology*, 33: 510–40.

Dovidio, J., and Gaertner, S. 2000. "Aversive Racism and Selection Decisions: 1989 and 1999." *Psychological Science*, 11: 319–23.

Dovidio, J., Kawakami, K., and Gaertner, S. 2002. "Implicit and explicit prejudice and interracial interaction." *Journal of Personality and Social Psychology*, 82: 62–68.

Dovidio, J., and Gaertner, S. 1986. "The aversive form of racism," in J. Dovidio and S. Gaertner (Eds.), *Prejudice, Discrimination and Racism*, (pp. 61–89). Academic Press.

Drake, S., & Cayton, H. 1945/1993. *Black Metropolis: A Study of Negro Life in a Northern City.* Chicago: University of Chicago Press.

Dreier, J. 1996. "Rational Preference: Decision Theory as a Theory of Practical Rationality." *Theory and Decision*, 40: 249–276.

Duneier, M. 2016. *Ghetto: The Invention of a Place, The History of an Idea.* New York: Farrar, Straus, and Giroux.

Dunning, D., & Cohen, G. 1992. "Egocentric definitions of traits and abilities in social judgment." *Journal of Personality and Social Psychology*, 63: 341–355.

Egan, L., Santos, L., and Bloom, P. 2007. "The origins of cognitive dissonance: Evidence from children and monkeys." *Psychological Science*, 11: 978–983.

Elster, J. 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Ernest-Jones, M., Nettle, D., and Bateson, M. 2011. "Effects of eye images on everyday cooperative behavior: A field experiment." *Evolution and Human Behavior*, 32: 172–178.

Evans, J. 2007. *Hypothetical thinking: Dual processes in reasoning and judgement.* Hove: Psychology Press.

Evans, J., & Frankish, K. (Eds.). 2009. *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.

Evans, J., and Over, D. 1996. *Rationality and reasoning.* Hove: Psychology Press.

Faranda, J., and Gaertner, S. 1979. "The effects of inadmissible evidence introduced by the prosecution and the defense, and the defendant's race on the verdicts by high and low authoritarians." Paper presented at the annual meeting of the *Eastern Psychological Association* (March), New York.

Farnham, S., Greenwald, A., and Banaji, M. 1999. "Implicit self-esteem." In D. Abrams & M. Hogg (Eds.), *Social identity and social cognition, Vol. 27*, (pp. 230–248). Malden, MA: Blackwell.

Faucher, L., & Machery, E. 2009. "Racism: Against Jorge Garcia's Moral and Psychological Monism." *Philosophy of the Social Sciences*, 39: 41–62.

Fazio, R. 2007. "Attitudes as Object-Evaluation Associations of Varying Strength." *Social Cognition*, 25: 603–37.

Fein, S., and Spencer, S. 1997. "Prejudice as self-image maintenance: Affirming the self through derogating others." *Journal of Personality and Social Psychology*, 73: 31–44.

Fein, S., Hoshino-Browne, E., Davies, P., and Spencer, S. 2003. "The role of self-image maintenance in stereotype activation and application." In S. Spencer, S. Fein, M. Zanna, & J. Olson (Eds.), *Motivated social perception: The Ontario Symposium*, Vol. 9, (pp. 21–44). Mahwah, NJ: Erlbaum.

Feltz, A. 2012. "Pereboom and Premises: Asking the Right Questions in the Experimental Philosophy of Free Will." *Consciousness and Cognition*, 22: 53–63.

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

Festinger, L, & Carlsmith, J. 1959. "Cognitive consequences of forced compliance." *Journal of Abnormal and Social Psychology*, 58: 203–210.

Fischer, G., & Hawkins, S. 1993. "Strategy compatibility, scale compatibility, and the prominence effect." *Journal of Experimental Psychology: Human Perception and Performance*, 19: 580–597.

Fischer, J. M. & Rivazza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility.* Cam-

bridge: Cambridge University Press.

Fischer, J. M. and Tognazzini, N. 2009. "The Truth about Tracing." *Noûs*, 43: 531–56.

Follenfant, A. & Ric, F. 2010. "Behavioral rebound following stereotype suppression." *European Journal of Social Psychology*, 40: 774–782.

Forscher, P., Lai, C., Axt, J., Ebersole, C., Herman, M., Devine, P., and Nosek, B. Manuscript. "A meta-analysis of change in implicit bias."

Frankfurt, H. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy*, 68: 5–20. Reprinted in *The Importance of What We Care About*, 1988, (pp. 11–25). Cambridge: Cambridge University Press.

Frankfurt, H. 1982. "The Importance of What We Care About." *Synthese*, 53: 257–272. Reprinted in *The Importance of What We Care About*, 1988, (pp. 80–95). Cambridge: Cambridge University Press.

Frankfurt, H. 1987. "Identification and Wholeheartedness." Reprinted in *The Importance of What We Care About*, 1988, (pp. 159–176). Cambridge: Cambridge University Press.

Frankfurt, H. 1999. "The Faintest Passion." In *Necessity, Volition, and Love*, (pp. 95–107). Cambridge: Cambridge University Press.

Frankfurt, H. 2002a. "Reply to T. M. Scanlon." In S. Buss & L. Overton (Eds.), *The Contours of Agency: Essays on Themes from Harry Frankfurt*, (pp. 184–188). Cambridge, MA: MIT Press.

Frankfurt, H. 2002b. "Reply to Barbara Herman." In S. Buss & L. Overton (Eds.), *The Contours of Agency: Essays on Themes from Harry Frankfurt*, (pp. 275–278). Cambridge, MA: MIT Press.

Frankfurt, H. 2006a. *The Reasons of Love*. Princeton, NJ: Princeton University Press.

Frankfurt, H. 2006b. *Taking Ourselves Seriously and Getting It Right*. Stanford, CA: Stanford University Press.

Freedman, D. 2012. "The perfected self." *The Atlantic*, June.

Friedman, M. 1986. "Autonomy and the Split-Level Self." *Southern Journal of Philosophy*, 24: 19–35.

Funder, D., & Ozer, D. 1983. "Behavior as a Function of the Situation." *Journal of Personality and Social Psychology*, 44: 107–112.

Galdi, S., Arcuri, L., and Gawronski, B. 2008. "Automatic Mental Associations Predict Future Choices of Undecided Decision-Makers." *Science*, 321: 1100-1102.

Gale, W., Iwry, J., John, D., and Walker, A. 2009. *Automatic: Changing The Way America Saves*. Washington, DC: Brookings Institution Press.

Gardner, W., Gabriel, S., and Lee, A. 1999. "'I' value freedom but 'we' value relationships: Self-construal priming mirrors cultural differences in judgment." *Psychological Science*, 10: 321–326.

Gardner, W., Gabriel, S., and Hochschild, L. 2002. "When you and I are 'we', you are no longer threatening: The role of self-expansion in social comparison processes." *Journal of Personality and Social Psychology*, 83: 239–251.

Gawronski, B., Geschke, D., and Banse, R. 2003. "Implicit bias in impression formation: associations influence the construal of individuating information." *European Journal of Social Psychology*, 33: 573–589.

Gawronski, B., Walther, E., and Blank, H. 2005. "Cognitive Consistency and the Formation of Interpersonal Attitudes: Cognitive Balance Affects the Encoding of Social Information." *Journal of Experimental Social Psychology*, 41: 618–626.

Gawronski, B., Bodenhausen, G., and Becker, A. 2007. "I like it, because I like myself: Associative self-anchoring and post-decisional change of implicit evaluations." *Journal of Experimental Social Psychology*, 43: 221–32.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., and Strack, F. 2008. "When 'just say no' is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation." *Journal of Experimental Social Psychology*, 44: 370–377.

Gawronski, B., and Bodenhausen, G. 2011. "The Associative-Propositional Evaluation Model: Theory, Evidence, and Open Questions." *Advances in Experimental Social Psychology*, 44: 59–127.

Gendler, T. 2008a. "Alief and Belief." *Journal of Philosophy*, 105: 634–63.

Gendler, T. 2008b. "Alief in Action (and Reaction)." *Mind & Language*, 23: 552–85.

Gendler, T. 2011. "On the Epistemic Costs of Implicit Bias." *Philosophical Studies*, 156: 33–63.

Gender, T. 2014. "The Third Horse: On Unendorsed Association and Human Behaviour." *Proceedings of the Aristotelian Society*, 88: 185–218.

Gerstenberg, T., & Lagnado, D. Forthcoming. "Attributing Responsibility: Actual and Counterfactual Worlds." *Oxford Studies in Experimental Philosophy*, Vol. 1.

Gigerenzer, G., Todd, P., and the ABC Research Group. 1999. *Simple Heuristics That Make Us Smart.* Oxford: Oxford University Press.

Gigerenzer, G. 2007. *Gut feelings: The intelligence of the unconscious.* New York: Penguin Press.

Gilbert, D. 1991. "How Mental Systems Believe." *American Psychologist*, 46: 107–19.

Gilboa, I., & Schmeidler, D. "Case-based decision theory." *Quarterly Journal of Economics*, 110: 605–639.

Gilles, S. 2011. *The tyranny of utility: Behavioural social science and the rise of paternalism.* Princeton: Princeton University Press.

Gilovich, T., Griffin, D., and Kahneman, D. 2002. *Heuristics and biases: The psychology of intuitive judgment.* Cambridge: Cambridge University Press.

Glaser, J., & Banaji, M. 1999. "When fair is foul and foul is fair: Reverse priming in automatic evaluation." *Journal of Personality and Social Psychology*, 77: 669–687.

Glaser, J. 2003. "Reverse priming: Implications for the (un)conditionality of automatic evaluation." In J. Musch and K. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion*, (pp. 87–108). Yahweh, NJ: Erlbaum.

Glaser, J., & Kihlstrom, J. 2006. "Compensatory automaticity: Unconscious volition is not an oxymoron." In R. Hassin, J. Uleman, and J. Bargh (Eds.), *The New Unconscious*. Oxford: Oxford University Press.

Glasgow, J. 2016. "Alienation and Responsibility." In M. Brownstein and J. Saul (Eds.), *Implicit Bias & Philosophy, Vol. 2*, (pp. 37–61). Oxford: Oxford University Press.

Gneezy, U., & Rustichini, A. 2000a. "A fine is a price." *Journal of Legal Studies*, 29: 1–17.

Gneezy, U., & Rustichini, A. 2000b. "Pay enough or don't pay at all." *Quarterly Journal of Economics*, 115: 791–810.

Goff, P., Jackson, M., Di Leone, B., Culotta, C., and DiTomasso, N. 2014. "The essence of innocence: Consequences of dehumanizing black children." *Journal of Personality and Social Psychology*, 106: 526–545.

Goldberger, J., Wheeler, G., and Sydenstrycker, E. 1920. "A study of the relation of family income and other economic factors to Pellagra incidence in seven cotton mill villages of South Carolina in 1916." *Public Health Reports*, 35: 2673–2714.

Goleman, D. 1995. *Emotional Intelligence*. New York: Bantam Books.

Gollwitzer, P., Bayer, U., and McCulloch, K. 2005. "The control of the unwanted." In R. Hassin, J. Uleman, & J. Bargh (Eds.), *The New Unconscious*, (pp. 485–515). Oxford: Oxford University Press.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science*, 293: 2105–8.

Greene, J., and Haidt, J. 2002. "How (and where) does moral judgment work?" *Trends in Cognitive Sciences*, 6: 517–523.

Greene, J., Nystrom, L., Engell, A., Darley, J., Cohen, J. 2004. "The neural bases of cognitive conflict and control in moral judgment." *Neuron*, 44: 389–400.

Greene, J., and Cohen, J. 2004. "For the law, neuroscience changes nothing and everything." *Philosophical Transactions of the Royal Society B, Biological Sciences*, 359:1775–1785.

Greene, J., Cushman, F., Stewart. L., Lowenberg, K., Nystrom, L., and Cohen, J. 2009. "Pushing moral buttons: The interaction between personal force and intention in moral judgment." *Cognition*, 111: 364–371.

Greene, J. 2013. *Moral Tribes: Emotion, Reason, and The Gap Between Us and Them.* New York: Penguin Press.

Greene, J. 2014. "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)science Matters for Ethics." *Ethics*, 124: 695–726.

Greenwald, A., and Banaji, M. 1995. "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes." *Psychological Review*, 102: 4–27.

Greenwald, A., Banaji, M., Rudman, L., Farnham, S., Nosek, B., and Mellott, D. 2002. A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept." *Psychological Review*, 109: 3–25.

Greenwald, A., Poehlman, T., Uhlmann, E., and Banaji, M. 2009. "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity." *Journal of Personality and Social Psychology*, 97: 17–41.

Grether, D., and Plott, C. 1979. "Economic Theory of Choice and the Preference-Reversal Phenomenon." *American Economic Review*, 69: 623–38.

Grüne-Yanoff, T. 2012. "Old wine in new casks: Libertarian paternalism still violates liberal principles." *Social Choice and Welfare*, 38: 635–645.

Haidt, J. 2001. "The Emotional Dog and its Rational Tale: A Social Intuitionist Approach to Moral Judgment." *Psychological Review*, 108: 814–34.

Haley, K., & Fessler, D. 2005. "Nobody's watching? Subtle cues affect generosity in an anonymous economic game." *Evolution and Human Behavior*, 26: 245–256.

Hall, C., Galvez, M., Sederbaum, I. 3024. "Assumptions about behavior and choice in response to public assistance: A behavioral decision analysis." *Policy Insights from the Behavioral and Brain Sciences*, 1: 137–143.

Hammond, K., Stewart, T., Brehmer, B., and Steinmann, D. 1975. "Social judgment theory." In M. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes*, (pp. 271–312). New York: Academic Press.

Hammond, K., McClellan, G., and Mumpower, J. 1980. *Human judgment and decision making: Theories, methods, and procedures.* New York: Praeger.

Harding, D. 2010. *Living the Drama: Community, Conflict, and Culture Among Inner-City Boys.* Chicago: Chicago University Press.

Harman, G. 1999. "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error." *Proceedings of the Aristotelian Society*, 99: 315–331.

Harman, E. 2007. "Discussion of Nomy Arpaly's *Unprincipled Virtue*." *Philosophical Studies*, 134: 433–439.

Hassin, R., Aarts, H., Eitam, B., Custers, R., and Kleiman, T. 2009. "Non-conscious goal pursuit and the effortful control of behavior." In E. Morsel, J. Bargh, and P. Gollwitzer (Eds.), *Oxford handbook of the psychology of action*, (pp. 549–568). Oxford: Oxford University Pres.

Hasson, U., and Glucksberg, S. 2006. "Does Negation Entail Affirmation? The Case of Negated Metaphors." *Journal of Pragmatics*, 38: 1015–32.

Hausman, D. 1992. *The Inexact and Separate Science of Economics.* Cambridge: Cambridge University

Press.

Hausman, D., & Welch, B. 2010. "Debate: To Nudge or Not to Nudge." *The Journal of Political Philosophy*, 18: 123–136.

Hausman, D. 2012. *Preference, Value, Choice, and Welfare.* Cambridge: Cambridge University Press.

Hawkins, S. 1994. "Information processing strategies in riskless preference reversals: The prominence effect." *Organizational Behavior and Human Decision Processes*, 59: 1–26.

Heider, F. 1958. *The psychology of interpersonal relations.* John Wiley & Sons.

Hitchcock, C. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *Journal of Philosophy*, 98: 273–299.

Hitchcock, C., & Woodward, J. 2003. "Explanatory Generalizations, Part II: Plumbing Explanatory Depth." *Noûs*, 37: 181–199.

Hitlin, S., and Vaisey, S. 2013. "The New Sociology of Morality." *Annual Review of Sociology*, 39: 51–68.

Hodson, G., Dovidio, J., and Gaertner, S. 2002. "Processes in racial discrimination: Differential weighting of conflicting information." *Personality and Social Psychology Bulletin*, 28: 460–71.

Holland, R., Wennekers, A., Bijlstra, G., Jongenelen, M., and van Knippenberg, A. 2009. "Self-symbols as implicit motivators." *Social Cognition*, 27: 579–600.

Holroyd, J. 2012. "Responsibility for Implicit Bias." *Journal of Social Science*, 43: 274–306.

Horowitz, R. 1983. *Honor and the American Dream.* New Brunswick, NJ: Rutgers University Press.

Hsee, C., Loewenstein, G., Blount, S., and Bazerman, M. 1999. "Preference reversals between joint and separate evaluation of options: a review and theoretical analysis." *Psychological Bulletin*, 125: 576–590.

Huber, J., Payne, J., and Puto, C. 1982. "Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis." *Journal of Consumer Research*, 9: 90–98.

Hume, D. 1738/1978. *A Treatise of Human Nature.* Oxford: Oxford University Press.

Hugenberg, K., & Bodenhausen, G. 2003. "Facing prejudice: Implicit prejudice and the perception of facial threat." *Psychological Science*, 14: 640–643.

Hursthouse, R. 1991. "Arational Actions." *Journal of Philosophy*, 88: 57–68.

Inbar, Y., Pizarro, D., Knobe, J., and Bloom, P. 2009. "Disgust sensitivity predicts intuitive disapproval of gays." *Emotion*, 9: 435–439.

Isen, A., & Levin, P. 1972. "Effect of feeling good on helping: Cookies and kindness." *Journal of Personality and Social Psychology*, 21: 384–388.

Ismael, J. 2013. "Causation, Free Will, and Naturalism." In H. Kincaid, J. Ladyman, and D. Ross (Eds.), *Scientific Metaphysics*, (pp. 208–235). Oxford: Oxford University Press.

Jacobs, R., & Campbell, D. 1961. "Transmission of an arbitrary social tradition." *Journal of Abnormal and Social Psychology*, 62: 649–658.

Jacobson, D. 2012. "Moral Dumbfounding and Moral Stupefaction." In M. Timmons (ed.), *Oxford Studies in Normative Ethics: Volume 2* (pp. 289–316). Oxford: Oxford University Press.

James, W. 1890. *The Principles of Psychology.* Cambridge, MA: Harvard University Press.

Jargowsky, P. 1997. *Poverty and Place: Ghettos, Barrios, and the American City.* New York: Russell Sage Foundation.

Jaworska, A. 2007. "Caring and Internality." *Philosophy and Phenomenological Research*, 74: 529–568.

Johnson, E., & Goldstein, D. 2003. "Do defaults save lives?" *Science*, 302: 1338–1339.

Johnson, J., Whitestone, E., Jackson, L., and Gatto, L. 1995. "Justice is still not colorblind: Differential racial effects of exposure to inadmissible evidence." *Personality and Social Psychology Bulletin*, 21: 893–8.

Jolls, C., & Sunstein, C. 2006a. "Debiasing Through Law." *Journal of Legal Studies*, 35: 199–241.

Jolls, C., & Sunstein, C. 2006b. "The Law of Implicit Bias." *California Law Review*, 94: 969–996.

Jones, J., Pelham, B., and Mirenberg, M. 2002. "Name letter preferences are not merely mere exposure: Implicit egotism as self-regulation." *Journal of Experimental Social Psychology*, 38: 170 –77.

Jost, J., Rudman, L., Blair, I., Carney, D., Dasgupta, N., Glaser, J., and Hardin, C. 2009. "The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore." *Research in Organizational Behavior*, 29: 39–69.

Kahneman, D. 1973. *Attention and effort.* Englewood Cliffs, NJ: Prentice Hall.

Kahneman, D., & Tversky, A. 1979. "Prospect theory." *Econometrica*, 47: 263–291.

Kahneman, D., Knetsch, J., and Thaler, R. 1991. "Anomalies: The endowment effect, loss aversion, and status quo bias." *Journal of Economic Perspectives*, 5: 193–206.

Kahneman, D., & Frederick, S. 2002. "Representativeness revisited: Attribute substitution in intuitive judgment." In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment*, (pp. 49–81). Cambridge: Cambridge University Press.

Kahneman, D. 2011. *Thinking, Fast and Slow.* New York: Farrar, Straus, and Giroux.

Kahneman, D., Slovic, P., and Tversky, A. (Eds.). 1982. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kang, J., & Banaji, M. 2006. "Fair Measures: A Behavioral Realist Revision of 'Affirmative Action'." *California Law Review*, 94: 1072–1075.

Kane, R. 1996. *The Significance of Free Will.* Oxford: Oxford University Press.

Kawakami, K., Dovidio, J., Moll, J., Hermsen, S. and Russin, A. 2000. "Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation." *Journal of Personality and Social Psychology*, 78: 871–888.

Kawakami, K., Dovidio, J., and van Kamp, S. 2007. "The impact of counterstereotypic training and related correction processes on the application of stereotypes." *Group Processes and Intergroup Relations*, 10: 139–156.

Kawakami, K., Phills, C., Steele, J., and Dovidio, J. 2005. "(Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors." *Journal of Personality and Social Psychology*, 92: 957–971.

Kelly, D., and Roedder, E. 2008. "Racial Cognition and The Ethics of Implicit Bias." *Philosophy Compass*, 3: 522–540.

Khoury, A. 2014. "Manipulation and Mitigation." *Philosophical Studies*, 168: 283–294.

Kirschenman, J., & Neckerman, K. 1991. "We'd Like to Hire Them, But..." In C. Jencks & P. Peterson (Eds.), *The Urban Underclass*, (pp. 203–232). Washington, D.C.: Brookings Institution.

Kitayama, S., and Karasawa, M. 1997. "Implicit self-esteem in Japan: Name letters and birthday numbers." *Personality and Social Psychology Bulletin*, 23: 736–742.

Knobe, J. 2003. "Intentional Action and Side Effects in Ordinary Language." *Analysis*, 63: 190–194.

Kolodny, N. 2005. "Why Be Rational?" *Mind*, 114: 509–563.

Koole, S., Dijksterhuis, A., and van Knippenberg, A. 2001. "What's in a name: Implicit self-esteem and the automatic self." *Journal of Personality and Social Psychology*, 80: 669–685.

Korsgaard, C. 1996. *The Sources of Normativity.* New York: Cambridge University Press.

Krysan, M., & Bader, M. 2009. "Racial blind spots: Black-white-Latino differences in community knowledge." *Social Problems*, 56: 677–701.

Kunda, Z. 1987. "Motivated inference: Self-serving generation and evaluation of causal theories." *Journal of Personality and Social Psychology*, 53: 37–54.

Kunda, Z. 1990. "The case for motivated reasoning." *Psychological Bulletin*, 108: 480–498.

Kunda, Z., & Spencer, S. 2003. "When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application." *Psychological Bulletin*, 129: 522–544.

Kurtz, S., & Saks, M. 1996. "The Transplant Paradox: Overwhelming Public Support for Organ Dona-

tion vs. Under-Supply of Organs: The Iowa Organ Procurement Study." *Journal of Corporation Law*, 21: 767–806.

Labov, W. 1972. *Language in the Inner City: Studies in the Black English Vernacular.* Philadelphia: University of Pennsylvania Press.

Labov, W. 2012. *Dialect Diversity in America: The Politics of Language Change.* Charlottesville: University of Virginia Press.

Lacetera, N., & Macis, M. 2010. "Do all material incentives for pro-social activities backfire? The response to cash and non-cash incentives for blood donations." *Journal of Economic Psychology*, 31: 738–748.

Lacetera, N., Macis, M., Slonim, R. 2012. "Will there be blood? Incentives and displacement effects in pro-social behavior." *American Journal of Economic Policy*, 4: 186–223.

Lagnado, D., & Channon, S. 2008. "Judgments of cause and blame: The influence of intentionality and foreseeability." *Cognition*, 108: 754–70.

Lagnado, D., Gerstenberg, T., and Zultan, R. 2013. "Causal responsibility and counterfactuals." *Cognitive Science*, 37: 1036–73.

Lai, C. et al. 2016. "Reducing implicit racial preferences: II. Intervention effectiveness across time." *Journal of Experimental Psychology*, 145: 1001–1016.

Leslie, S. Forthcoming. "The Original Sin of Cognition: Fear, Prejudice, and Generalization." *Journal of Philosophy.*

Levitt, S., & Venkatesh, S. 2000. "An Economic Analysis of a Drug-Selling Gang's Finances." *The Quarterly Journal of Economics*, 115: 755–789.

Levy, N. (Ed.) 2013. *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience.* Oxford: Oxford University Press.

Levy, N. 2014a. "Consciousness, Implicit Attitudes, and Moral Responsibility." *Noûs*, 48: 21–40.

Levy, N. 2014b. *Consciousness & Moral Responsibility.* Oxford: Oxford University Press.

Levy, N. 2015. "Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements." *Noûs*, 49: 800–823.

Lewis, O. 1959. *Five Families: Mexican Case Studies in the Culture of Poverty.* New York: Basic Books.

Lewis, O. 1961. *The Children of Sanchez.* New York: Random House.

Lewis, O. 1966. *La Vida: A Puerto Rican Family in the Culture of Poverty—San Juan and New York.* New York: Random House.

Lewis, O. 1968. "The Culture of Poverty." In D. Moynihan (ed.), *On Understanding Poverty: Perspective from the Social Sciences*, (pp. 187–200). New York: Basic Books.

Libet, B. 1985. "Unconscious cerebral initiative and the role of conscious will in voluntary action." *Be-*

*havioral and Brain Sciences*, 8: 529–566.

Lichtenstein, S., and Slovic, P. 1971. "Reversals of Preferences between Bids and Choices in Gambling Decisions." *Journal of Experimental Psychology*, 89: 46–55.

Lichtenstein, S., and Slovic, P. 1973. "Response-Induced Reversals of Preference in Gambling: An Extended Replication in Las Vegas." *Journal of Experimental Psychology*, 101: 16–20.

Lichtenstein, S., and Slovic, P. (Eds.). 2006. *The Construction of Preference*. Cambridge: Cambridge University Press.

Lieberman, M., Ochsner, K., Gilbert, D., and Schacter, D. 2001. "Do amnesiacs exhibit cognitive dissonance reduction? The role of explicit memory and attention in attitude change." *Psychological Science*, 2: 135–140.

Liu, B., & Ditto, P. 2013. "What dilemma? Moral evaluation shapes factual belief." *Social Psychological and Personality Science*, 4: 316–323.

Locke, A. 1925. "Enter the New Negro," in A. Locke (Ed.), *The New Negro*, *Survey Graphic*, March 1925.

Loewenstein, G., & Issacharoff, S. 1994. "Source dependence in the valuation of objects." *Journal of Behavioral Decision Making*, 7: 157–168.

Logan, J., Stults, B., and Farley, R. 2004. "Segregation of minorities in the metropolis: Two decades of change." *Demography*, 41: 1–22.

Lombrozo, T. 2010. "Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions." *Cognitive Psychology*, 61: 303–332.

Lowery, B., Hardin, C., Sinclair, S. 2001. "Social Influence Effects on Automatic Racial Prejudice" *Journal of Personality and Social Psychology*, 81: 842–855.

Luce, D. 1959. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Machery, E., Faucher, L., and Kelly, D. 2010, "On the Alleged Inadequacy of Psychological Explanations of Racism." *The Monist*, 93: 228–255.

Mackie, D., Hamilton, D., Susskind, J., and Rosseli, F. 1996. "Social psychological foundations of stereotype formation." In C. Macra, C. Stangor, and M. Hewstone (Eds.), *Stereotypes and stereotyping*, (pp. 41–78). New York: Guilford Press.

Madva, A. 2016. "Virtue, Social Knowledge, and Implicit Bias." In M. Brownstein & J. Saul (Eds.), *Implicit Bias & Philosophy: Volume I, Metaphysics and Epistemology*. Oxford: Oxford University Press.

Mandelbaum, E. 2013. "Against Alief." *Philosophical Studies*, 165: 197–211.

Mandelbaum, E. Forthcoming. "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias." *Noûs*.

Mani, A., Mullainathan, S., Shafir, E., and Zhao, J. 2013. "Poverty impedes cognitive function." *Science*, 341: 976–980.

Marien, H., Custers, R., Hassin, R., and Aarts, H. 2012. "Unconscious goal activation and the hijacking of the executive function." *Journal of Personality and Social Psychology*, 103: 399–415.

Massey, D. 1990. "American Apartheid: Segregation and the Making of the Underclass." *American Journal of Sociology*, 96: 329–357.

Massey, D., & Eggers, M. 1990. "The Ecology of Inequality: Minorities and the Concentration of Poverty, 1970–1980." *American Journal of Sociology*, 95: 1153–1188.

Massey, D., & Denton, N. 1993. *American Apartheid: Segregation and the Making of the Underclass.* Cambridge, MA: Harvard University Press.

Massey, D., & Sampson, R. 2009. "Moynihan Redux: Legacies and Lessons." *Annals of the American Academy of Political and Social Science*, 621: 6–27.

Massey, D., Albright, L., Casciano, R., Derickson, E., and Kinsey, D. 2013. *Climbing Mount Laurel: The Struggle for Affordable Housing and Social Mobility in an American Suburb.* Princeton: Princeton University Press.

McClennen, E. 1990. *Rationality and Dynamic Choice: Foundational Explorations.* Cambridge: Cambridge University Press.

Medvec, V., Madey, S., and Gilovich, T. 1995. "When less is more: Counterfactual thinking and satisfaction among Olympic medalists." *Journal of Personality and Social Psychology*, 69: 603–610.

Mele, A. 2009. *Effective Intentions: The Power of Conscious Will.* Oxford: Oxford University Press.

Mele, A. 2013. *A Dialogue on Free Will and Science.* Oxford: Oxford University Press.

Mellström, C., & Johannesson, M. 2008. "Crowding out in blood donation: Was Titmuss right?" *Journal of the European Economic Association*, 6: 845–863.

Mendoza, S., Gollwitzer, P., and Amodio, D. 2010. "Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions." *Personality and Social Psychology Bulletin*, 36: 512–523.

Milgram, S. 1970. "The experience of living in cities." *Science*, 167: 1461–1468.

Milgram, S. 1974. *Obedience to authority.* New York: Harper and Row.

Montgomery, H. 1983. "Decision rules and the search for a dominance structure: Towards a process model of decision making." In P. Humphreys, O. Svenson, and A. Vari (Eds.), *Analysing and aiding decision processes*, (pp. 343–369). Amsterdam: North-Holland and Budapest: Akademiai Kiado.

Montgomery, H. 1989. "From cognition to action: The search for dominance in decision making." In H. Montgomery & O. Svenson (Eds.), *Process and structure in human decision making*, (pp. 23–49). Chichester, England: Wiley.

Montgomery, H. 1993. "The search for a dominance structure in decision making: Examining the evidence." In G. Klein, J. Orasanu, R. Calderwood, and C. Zsambok (Eds.), *Decision making in action: Models*

*and methods*, (pp. 182–187). Norwood, NJ: Ablex.

Montgomery, H., & Svenson, O. 1983. "A think aloud study of dominance structuring in decision processes." In R. Tietz (Ed.), *Aspiration levels in bargaining and economic decision making*, (pp. 383–399). Berlin: Springer-Verlag.

Moody-Adams, M. 1994. "Culture, Responsibility, and Affected Ignorance." *Ethics*, 104: 291–309.

Morewedge, C., Shu, L., Gilbert, D., and Wilson, T. 2009. "Bad riddance or good rubbish? Ownership and not loss aversion causes the endowment effect." *Journal of Experimental Social Psychology*, 45: 947–951.

Moskowitz, G., Gollwitzer, P., Wasel, W., and Schaal, B. 1999. "Preconscious control of stereotype activation through chronic egalitarian goals." *Journal of Personality and Social Psychology*, 77: 167–184.

Moskowitz, G., and Li, P. 2011. "Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control." *Journal of Experimental Social Psychology*, 47: 103–116.

Mullainathan, S., & Shafir, E. 2013. *Scarcity: The New Science of Having Less and How it Defines Our Lives.* New York: Picador.

Murray, D., & Nahmias, E. 2014. "Explaining Away Incompatibilist Intuitions." *Philosophy and Phenomenological Research*, 88: 434–467.

Murray, D. 2015. "Situationism, Going Mental, and Modal *Akrasia*." *Philosophical Studies*, 172: 711–36.

Murray, D., and Lombrozo, T. 2017. "Effects of Manipulation on Attributions of Causation, Free Will, and Moral Responsibility." *Cognitive Science*, 41: 447-481.

Murray, D., & Buchak, L. (Under Review). "Risk and Motivation: Why Reason Alone Is Not Enough to Determine Rational Action."

Murray, D. (Forthcoming). "Character and Situationism." In M. Vargas and J. Doris (Eds.), *Oxford Handbook of Moral Psychology.* Oxford: Oxford University Press.

Myrdal, G. 1944. *An American Dilemma: The Negro Problem and Modern Democracy.* New York and London: Harper and Brothers.

Nahmias, E. 2007. "Autonomous Agency and Social Psychology." In M. Marraffa, M. Caro, and F. Ferretti (Eds.), *Cartographies of the Mind: Philosophy and Psychology in Intersection*, (pp. 169–185). Springer.

Nahmias, E. 2011. "Intuitions about Free Will, Determinism, and Bypassing." In R. Kane (ed.), *Oxford Handbook of Free Will* (2nd edition), (pp. 555–576). New York: Oxford University Press.

Nahmias, E. 2014. "Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences." In W. Sinnott-Armstrong (ed.), *Moral Psychology, Vol. 4: Freedom and Responsibility*, (pp. 1-25). Cambridge, MA: MIT Press.

Nahmias, E., Coates, D., and Kvaran, T. 2007. "Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions." *Midwest Studies in Philosophy*, 31: 214–242.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. 2005. "Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility." *Philosophical Psychology*, 18: 561–584.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. 2006. "Is Incompatibilism Intuitive?" *Philosophy and Phenomenological Research*, 73: 28–53.

Nahmias, E. & Murray, D. 2010. "Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions." In J. Aguilar, A. Buckareff, and K. Frankish (Eds.), *New Waves in Philosophy of Action* (pp. 189–216). Palgrave-Macmillan.

Navon, D. 1984. "Resources—A theoretical soup stone?" *Psychological Review*, 91: 216–234.

Nelkin, D. 2005. "Freedom, Responsibility, and the Challenge of Situationism." *Midwest Studies in Philosophy*, 29: 181–206.

Nelkin, D. Forthcoming. "Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness." *Noûs*.

Newman, G., Diesendruck, G., and Bloom, P. 2011. "Celebrity contagion and the value of objects." *The Journal of Consumer Research*, 38: 215–28.

Nichols, S. & Knobe, J. 2007. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Noûs*, 41: 663–685.

Nichols, S. 2011. "Experimental philosophy and the problem of free will." *Science*, 331: 1401–1403.

Nisbett, R., & Wilson, T. 1977. "Telling more than we can know: Verbal reports on mental processes." *Psychological Review*, 84: 231–59.

Nisbett, R., & Cohen, D. 1996. *Culture of honor: The psychology of violence in the south.* Boulder, CO: Westview Press.

Norton, M., Vandello, J., and Darley, J. 2004. "Casuistry and social category bias." *Journal of Personality and Social Psychology*, 87: 817–831.

Nosek, B. 2005. "Moderators of the relationship between implicit and explicit evaluation." *Journal of Experimental Psychology: General*, 134: 565–584.

Nosek, B. 2007. "Implicit-explicit relations." *Current Directions in Psychological Science*, 16: 65–69.

Nuttin, J. 1985. "Narcissism beyond gestalt and awareness: The name letter effect." *European Journal of Social Psychology*, 15: 353–361.

Nuttin, J. 1987. "Affective consequences of mere ownership: The name letter effect in twelve European languages." *European Journal of Social Psychology*, 17: 381–402.

Olin, L., and Doris, J. 2014. "Vicious Minds: Virtue Epistemology, Cognition, and Skepticism." *Philosophical Studies*, 168: 665–692.

Olsson, A., Ebert, J., Banaji, M., and Phelps, E. 2005. "The role of social groups in the persistence of

learned fear." *Science*, 309: 785–787.

Oswald, F., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P. 2013. "Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies." *Journal of Personality and Social Psychology*, 105: 171–192.

Otten, S. 2003. ""Me and us" or "us and them"? The self as a heuristic for defining minimal ingroups." *European Review of Social Psychology*, 13, 1–33.

Oyserman, D., and Lee, S. 2008. "Does culture influence what and how we think? Effects of priming individualism and collectivism." *Psychological Bulletin*, 134: 311–342.

Park, S., Glaser, J., and Knowles, E. 2008. "Implicit motivation to control prejudice moderates the effect of cognitive depletion on unintended discrimination." *Social Cognition*, 26: 401–419.

Pashler, H. 1998. *The psychology of attention.* Cambridge, MA: MIT Press.

Payne, B. 2001. "Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon." *Journal of Personality and Social Psychology*, 81: 181–192.

Payne, B., Cheng, C., Govorun, O., Stewart, B. 2005. "An Inkblot for Attitudes: Affect Misattribution as Implicit Measurement." *Journal of Personality and Social Psychology*, 89: 277–93.

Payne, B. 2006. "Weapon Bias: Split Second Decisions and Unintended Stereotyping." *Current Directions in Psychological Science*, 15: 287–91.

Payne, B., Krosnick, J., Pasek, J., Lelkes, Y., Akhtar, O., and Tompson, T. 2010. "Implicit and Explicit Prejudice in the 2008 American Presidential Election." *Journal of Experimental Social Psychology*, 46: 367–74.

Payne, J., Bettman, J., and Johnson, E. 1992. "Behavioral decision research: a constructive process perspective." *Annual Review of Psychology*, 43: 87–131.

Pearl, J. 2000. *Causality.* New York: Cambridge University Press.

Pelham, B., Mirenberg, M., and Jones, J. 2002. "Why Susie Sells Seashells by the Seashore: Implicit Egotism and Major Life Decisions." *Journal of Personality and Social Psychology*, 82: 469–87.

Pelham, B., Carvallo, M., and Jones, J. 2005. "Implicit Egotism." *Current Directions in Psychological Science*, 14: 106–110.

Pennington, N., & Hastie, R. 1988. "Explanation-based decision making: Effects of memory structure on judgment." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14: 521–533.

Pereboom, D. 2001. *Living Without Free Will.* Cambridge: Cambridge University Press.

Pereboom, D. 2014. *Free Will, Agency, and Meaning in Life.* New York: Oxford University Press.

Perkins, A., & Forehand, M. 2012. "Implicit self-referencing: The effect of nonvolitional self-association on brand and product attitude." *Journal of Consumer Research*, 39: 142–156.

Pettit, P. 1991. "Decision Theory and Folk Psychology." In M. Bacharach and S. Hurley (Eds.), *Foundations of Decision Theory: Issues and Advances*, (pp. 147–75). Oxford: Blackwell.

Pettit, P. 1996. "Freedom as Antipower." *Ethics*, 106: 576–604.

Phills, C., Kawakami, K., Tabi, E., Nadolny, D., and Inzlicht, M. 2011. "Mind the gap: Increasing the associations between the self and blacks with approach behaviors." *Journal of Personality and Social Psychology*, 100: 197–210.

Phillips, J., and Shaw, A. 2014. "Manipulating Morality: Third?Party Intentions Alter Moral Judgments by Changing Causal Reasoning." *Cognitive Science*, 38: 1320-1347.

Pockett, S. & Purdy, S. 2010. "Are voluntary movements initiated preconsciously? The relationships between readiness potentials, urges, and decisions." In W. Sinnott-Armstrong & L. Nadel (Eds.), *Conscious Will and Responsibility*, (pp. 34–46). Oxford: Oxford University Press.

Poundstone, W. 2008. *Gaming the Vote: Why Elections Aren't Fair (and What We Can Do About It).* New York: Hill and Wang.

Prestwich, A., Perugini, M., Hurling, R., and Richetin, J. 2010. "Using the Self to Change Implicit Attitudes." *European Journal of Social Psychology*, 40: 61–71.

Quek, B., and Ortony, A. 2011. "Modeling underlying mechanisms of the Implicit Association Test." *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 1330–1335.

Quillian, L. 1999. "Migration Patterns and the Growth of High-Poverty Neighborhoods, 1970–1990." *American Journal of Sociology*, 105: 1–37.

Quillian, L. 2012. "Segregation and poverty concentration: The role of three segregations." *American Sociological Review*, 77: 354–379.

Railton, P. 2009. "Practical Competence and Fluent Agency." In D. Sobel and S. Wall (Eds.), *Reasons for Action* (pp. 81–115). Cambridge: Cambridge University Press.

Railton, P. 2011. "Two Cheers for Virtue: Or, might Virtue be Habit Forming?" In M. Timmons (ed.), *Oxford Studies in Normative Ethics, Vol 1.* New York: Oxford University Press.

Railton, P. 2014. "The Affective Dog and Its Rational Tale: Intuition and Attunement." *Ethics*, 124: 813–859.

Rebonato, R. 2012. *Taking liberties: A critical examination of libertarian paternalism.* New York: Palgrave Macmillan.

Reiss, J. 2013. *Philosophy of Economics: A Contemporary Introduction.* New York: Routledge.

Rigdon, M., Ishii, K., Wantabe, M., and Kitayama, S. 2009. "Minimal social cues in the dictator game." *Journal of Economic Psychology*, 30: 358–367.

Risjord, M. 2014. *Philosophy of Social Science: A Contemporary Introduction.* New York: Rutledge.

Roskies, A. 2012. "Don't Panic: Self-Authorship without Obscure Metaphysics," *Philosophical Perspectives*, 26: 323–342.

Ross, L., & Nisbett, R. 1991. *The person and the situation: Perspectives of social psychology.* Philadelphia: Temple University Press.

Rozin, P., Millman, L., and Nemeroff, C. 1986. "Operation of the laws of sympathetic magic" in disgust and other domains." *Journal of Personality and Social Psychology*, 50: 703–12.

Rozin, P., Markith, M., and Ross, B. 1990. "The sympathetic magical law of similarity, nominal realism, and neglect of negatives in response to negative labels. *Psychological Science*, 1: 383–84.

Russo, J., Medvec, V., and Meloy, M. 1996. "The distortion of information during decisions." *Organizational Behavior and Human Decision Processes*, 66: 102–110.

Russo, J., Meloy, M., Medvec, V. 1998. "Predecisional distortion of product information." *Journal of Marketing Research*, 35: 438–452.

Russo, J., Meloy, M., Wilks, T. 2000. "Predecisional distortion of information by auditors and salespersons." *Management Science*, 46: 13–27.

Russo, J., Carlson, K., and Meloy, M. 2006. "Choosing an Inferior Alternative." *Psychological Science*, 17: 899–904.

Rydell, R., and McConnell, A. 2006. "Understanding Implicit and Explicit Attitude Change: A Systems of Reasoning Analysis." *Journal of Personality and Social Psychology*, 91: 995–1008.

Sacerdote, B. 2001. "Peer effects with random assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics*, 116: 681–704.

Salganik, M., Dodds, P., and Watts, D. 2006. "Experimental study of inequality and unpredictability in an artificial cultural market." *Science*, 311: 854–856.

Sampson, R., & Sharkey, P. 2008. "Neighborhood selection and the social reproduction of concentrated racial inequality." *Demography*, 45: 1–29.

Sampson, R. 2012. *Great American City: Chicago and the Enduring Neighborhood Effect.* Chicago: Chicago University Press.

Sánchez-Jankowski, M. 1991. *Islands in the Street: Gangs and American Urban Society.* Berkeley, CA: University of California Press.

Sánchez-Jankowski, M. 2008. *Cracks in the Pavement: Social Change and Resilience in Poor Neighborhoods.* Berkeley, CA: University of California Press.

Sartre, J. 1956. "Existentialism is a Humanism." In W. Kaufmann (ed.), *Existentialism from Dostoevsky to Sartre.* New York: Meridian/Penguin.

Saugstad, P., and Schioldborg, P. 1966. "Value and size perception." *Scandinavian Journal of Psychol-*

*ogy*, 7: 102–114.

Saul, J. Forthcoming. "Implicit Bias, Stereotype Threat and Women in Philosophy." In F. Jenkins and K. Hutchison (Eds.), *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press.

Savage, L. 1954/1972. *The Foundations of Statistics.* New York: Dover.

Schelling, T. 1978. *Micromotives and Macrobehavior.* New York: Norton.

Schimel, J., Simon, L., Greenberg, J., Pyszczynski, T., Solomon, S., Waxmonsky, J., and Arndt, J. 1999. "Stereotypes and terror management: Evidence that mortality salience enhances stereotypic thinking and preferences." *Journal of Personality and Social Psychology*, 77: 905–926.

Schnall, S., Haidt, J., Clore, G., and Jordan, A. 2008. "Disgust as embodied moral judgment." *Personality and Social Psychology Bulletin*, 34: 1096–1109.

Sechrist, G., & Stengor, C. 2001. "Perceived consensus influences intergroup behavior and stereotype accessibility." *Journal of Personality and Social Psychology*, 80: 645–654.

Sedikides, C. 1992. "Mood as a determinant of attentional focus." *Cognition and Emotion*, 6: 129–148.

Sedikides, C., Ariely, D., and Olsen, N. 1999. "Contextual and procedural determinants of partner selection: On asymmetric dominance and prominence." *Social Cognition*, 17: 118–139.

Shafir, E. 1993. "Choosing versus rejecting: why some options are both better and worse than others." *Memory & Cognition*, 21: 546–556.

Shafir, E. 2016. "Manipulated as a way of life." *Journal of Marketing Behavior*, 1: 245–260.

Shafir, E., & Tversky, A. 1992. "Thinking through uncertainty: non consequential reasoning and choice." *Cognitive Psychology*, 24: 449–474.

Shafir, E., Simonson, I., and Tversky, A. 1993. "Reason-based choice." *Cognition*, 49: 11–36.

Shah, A., Mullainathan, S., and Shafir, E. 2012. "Some consequences of having too little." *Science*, 338: 682–685.

Shampanier, K., Mazar, N, and Ariely, D. 2007. "Zero as a special price: The true value of free products." *Marketing Science*, 26: 742–757.

Sharkey, P. 2013. *Stuck in Place: Urban Neighborhoods and the End of Progress Toward Racial Equality.* Chicago: Chicago University Press.

Sharot T., Fleming, S., Yu, X., Koster, R., Dolan, R. 2012. "Is choice-induced preference change long lasting?" *Psychological Science*, 23: 1123–1129.

Shaw, C., & McKay, H. 1942. *Juvenile Delinquency and Urban Areas.* Chicago: University of Chicago Press.

Shelby, T. 2007. "Justice, Deviance, and the Dark Ghetto." *Philosophy & Public Affairs*, 35: 126–160.

Shelby, T. 2014. "Integration, Inequality, and Imperatives of Justice: A Review Essay." *Philosophy & Public Affairs*, 42: 253–285.

Shelby, T. 2016. *Dark Ghettos.* Cambridge, MA: Harvard University Press.

Sherif, M. 1937. "An Experimental Appraoch to the Study of Attitudes." *Sociometry*, 1: 90–98.

Shiffrin, S. 2000. "Paternalism, Unconscionability Doctrine, and Accomodation." *Philosophy & Public Affairs*, 29: 205–250.

Shiller, R. 2000. *Irrational Exuberance.* Princeton: Princeton University Press.

Shiller, R. 2008. *The Subprime Solution.* Princeton: Princeton University Press.

Shoemaker, D. 2011. "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics*, 121: 602–632.

Simon, D. 2004. "A third view of the black box: Cognitive coherence in legal decision making." *University of Chicago Law Review*, 71: 511–586.

Simon, D., Krawczyk, D. C., & Holyoak, K. J. 2004. "Construction of preferences by constraint satisfaction." *Psychological Science*, 15: 331–336.

Simonson, I. 1989. "Choice based on reasons: the case of attraction and compromise effects." *Journal of Consumer Research*, 16: 158–174.

Simonson, I., & Tversky, A. 1992. "Choice in context: tradeoff contrast and extremeness aversion." *Journal of Marketing Research*, 29: 281–295.

Simonson, I., Nowlis, S. and Simonson, Y. 1993. "The effect of irrelevant preference arguments on consumer choice." *Journal of Consumer Psychology*, 2: 287–306.

Simonson, I., Carmon, Z., and O'Curry, S. 1994. "Experimental evidence on the negative effect of product features and sales promotions on brand choice." *Marketing Science*, 13: 23–40.

Sinclair, L., & Kunda, Z. 1999. "Reactions to a black professional: Motivated inhibition and activation of conflicting stereotypes." *Journal of Personality and Social Psychology*, 77: 885–904.

Sinclair, L., & Kunda, Z. 2000. "Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me." *Personality and Social Psychology Bulletin*, 26: 1329–1342.

Sloman, S. 1996. "The empirical case for two systems of reasoning." *Psychological Bulletin*, 119: 3–22.

Slovic, P., and Lichtenstein, S. 1968. "Relative Importance of Probabilities and Payoffs in Risk Taking." *Journal of Experimental Psychology Monograph*, 78 (part 2): 1–18.

Slovic, P. 1975. "Choice between equally valued alternatives." *Journal of Experimental Psychology: Human Perception and Performance*, 1: 280–287.

Slovic, P. 1990. "Choice." In D. Osherson, N. Block, S. Kosslyn, and E. Smith (Eds.), *An invitation to cognitive science, Vol. 3*, (pp. 89–116). Cambridge, MA: MIT Press.

Slovic, P. 1995. "The construction of preference." *American Psychologist*, 50: 364–71.

Smith, A. 2005. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics*, 115: 236–71.

Smith, A. 2012. "Attributability, Answerability, and Accountability: In Defense of a Unified Account." *Ethics*, 122: 575–89.

Smith, H. 2015. "Dual-Process Theory and Moral Responsibility." In M. McKenna, A. Smith, and R. Clarke, (Eds.), *The Nature of Moral Responsibility*, (pp. 175208). Oxford: Oxford University Press.

Smith, M. 1998. "The Possibility of Philosophy of Action." In J. Bransen & Cuypers, S. (Eds.), *Human Action, Deliberation and Causation*, (pp. 17–41). Netherlands: Kluwer Academic Publishers.

Son Hing, L., Chung-Yan, G., Hamilton, L., and Zanna, M. 2008. "A two-dimensional model that employs explicit and implicit attitudes to characterize prejudice." *Journal of Personality and Social Psychology*, 94: 971–87.

Soon, C., Brass, M., Heinze, H., & Haynes, J. 2008. "Unconscious determinants of free decisions in the human brain." *Nature Neuroscience*, 11: 543–545.

Sowell, T. 2005. *Black Rednecks and White Liberals.* San Francisco: Encounter Books.

Spencer, S., Fein, S., Wolfe, C., Fong, C., and Dunn, M. 1998. "Automatic activation of stereotypes: The role of self-image threat." *Personality and Social Psychology Bulletin*, 24: 1139–1152.

Spirtes, P., Glymour, C., and Scheines, R. 1993. *Causation, Prediction and Search.* New York: Springer-Verlag.

Sripada, C. 2012. "What Makes a Manipulated Agent Unfree?" *Philosophy and Phenomenological Research*, 85: 563593.

Sripada, C. 2016. "Self-Expression: A Deep Self Theory of Moral Responsibility." *Philosophical Studies*, 173: 1203–1232.

Sripada, C. Forthcoming. "Free Will and the Construction of Options." *Philosophical Studies.*

Stanovich, K. 2004. *The robot's rebellion: Finding meaning in the age of Darwin.* Chicago: University of Chicago Press.

Stanovich, K. 2010. *Rationality and the reflective mind.* Oxford: Oxford University Press.

Steele, C. 2010. *Whistling Vivaldi: How stereotypes affect us and what we can do.* New York: W. W. Norton & Co.

Stewart, B., & Payne, B. 2008. "Brining automatic stereotyping under control: Implementation intentions as efficient means of thought control." *Personality and Social Psychology Bulletin*, 34: 1332–1345.

Strawson, P. 1962 . "Freedom and Resentment." *Proceedings of the British Academy*, 48: 1–25. Reprinted in G. Watson (ed.), *Free Will*, 2nd ed. (pp. 72–93). New York : Oxford University Press.

Suhler, C., & Churchland, P. 2009. "Control: Conscious and otherwise." *Trends in Cognitive Sciences*, 13: 341–347.

Sunstein, C. 2003. *Why Societies Need Dissent*. Cambridge: Harvard University Press.

Sunstein, C., Schkade, D., Ellman, L., and Sawicki, A. 2006. *Are Judges Political?* Washington, DC: Brookings Institution Press.

Sunstein, C. 2013. *Simpler: The Future of Government*. New York: Simon & Schuster.

Sunstein, C. 2014. *Why Nudge?: The Politics of Libertarian Paternalism*. Yale: Yale University Press.

Sunstein, C. 2015. "Ethical Nudging." *Yale Journal on Regulation*.

Sunstein, C. Forthcoming. *The Ethical State*.

Sunstein, C. Manuscript. "Which nudges do people like? A national survey."

Swartz, S. 2010. "'Moral ecology' and 'moral capital': tools towards a sociology of moral education from a South African Ethnography." *Journal of Moral Education*, 39: 305–327.

Thaler, R., & Sunstein, C. 2003a. "Libertarian paternalism." *American Economic Review*, 93: 175–179.

Thaler, R., & Sunstein, C. 2003b. "Libertarian paternalism is not an oxymoron." *University of Chicago Law Review*, 70: 1159–1202.

Thaler, R., & Sunstein, C. 2006. "Preferences, paternalism, and liberty." In S. Olsaretti (ed.), *Preferences and Well-Being* (pp. 233–264). Cambridge: Cambridge University Press.

Thaler, R., & Sunstein, C. 2008. *Nudge*. New Haven, CT: Yale University Press.

Thomas, W., & Znaniecki, F. 1918–1920. *The Polish Peasant in Europe and America*, 5 vols. Boston: Brager.

Thompson, C. 2007. "Desktop Orb Could Reform Energy Hogs." *Wired*, 15, July 24.

Tietje, B., & Brunel, F. 2005. "Towards a unified theory of implicit consumer brand cognitions." In F. Kardes, P. Herr, and J. Natel (Eds.), *Applying social cognition to consumer-focused strategy*, (pp. 135–153). Mahwah, NJ: Erlbaum.

Titmuss, R. 1970. *The Gift Relationship: From Human Blood to Social Policy*. London: Allen & Unwin.

Townsend, S. 2013. "The Ministry of Mind Control." *Property Week*, May 31.

Trevena, J. & Miller, J. 2009. "Brain preparation before a voluntary action: Evidence against unconscious movement preparation." *Consciousness and Cognition*, 19: 447–456.

Tversky, A., & Kahneman, D. 1974. "Judgments under uncertainty: Heuristics and biases." *Science*, 185: 1124–1131.

Tversky, A., & Kahneman, D. 1981. "The framing of decisions and the psychology of choice." *Science*, 211: 453–58.

Tversky, A., & Kahneman, D. 1983. "Extension versus intuitive reasoning: The conjunction fallacy in probability judgment." *Psychological Review*, 90: 293–315.

Tversky, A., & Kahneman, D. 1986. "Rational choice and the framing of decisions." *Journal of Business*, 59: S251–S278.

Tversky, A., Sattath, S., and Slovic, P. 1988. "Contingent weighting in judgment and choice." *Psychological Review*, 95: 371–384.

Tversky, A., Slovic, P., and Kahneman, D. 1990. "The causes of preference reversal." *American Economic Review*, 80: 204–17.

Tversky, A., & Thaler, R. 1990. "Preference reversals." *Journal of Economic Perspectives*, 4: 201–11.

Tversky, A., & Shafir, E. 1992a. "The disjunction effect in choice under uncertainty." *Psychological Science*, 3: 305–309.

Tversky, A., & Shafir, E. 1992b. "Choice under conflict: the dynamics of deferred decision." *Psychological Science*, 3: 358–361.

Tversky, A., & Simonson, I. 1993. "Context-dependent preferences." *Management Science*, 39: 1179–1189.

Twain, M. 1876. *The Adventures of Tom Sawyer*. Hartford, CT: American Publishing Co.

Twain, M. 1885. *Adventures of Huckleberry Finn*. New York: Charles L. Webster & Co.

Uhlmann, E., and Cohen, G. 2005. "Constructed criteria: Redefining merit to justify discrimination." *Psychological Science*, 16: 474–80.

Uhlmann, E., Pizarro, D., and Ditto, P. 2009. "The motivated use of moral principles." *Judgment and Decision Making*, 4: 476–91.

Uhlmann, E., Brescoll, V., and Machery, E. 2010. "The motives underlying stereotype-based discrimination against members of stigmatized groups." *Social Justice Research*, 23: 1–16.

Ullman-Margalit, E., & Morgenbesser, S. 1977. "Picking and Choosing." *Social Research*, 44: 758–759.

Vaisey, S. 2009. "Motivation and Justification: A Dual-Process Model of Culture in Action." *American Journal of Sociology*, 114: 1675–1715.

Valian, V. 2005, "Beyond Gender Schemas: Improving the Advancement of Women in Academia." *Hypatia*, 20: 198–213.

Vargas, M. 2005. "The Trouble with Tracing." *Midwest Studies in Philosophy*, 29: 269–91.

Vargas, M. 2006. "On the Importance of History for Responsible Agency." *Philosophical Studies*, 127: 351–382.

Vargas, M. 2013a. *Building Better Beings: A Theory of Moral Responsibility*. New York: Oxford University Press.

Vargas, M. 2013b. "Situationism and Responsibility: Free Will in Fragments." In T. Vierkant, J. Kiverstein, and A. Clark (Eds.), *Decomposing the Will*, (pp. 400–416). New York: Oxford University Press.

Velleman, J. D. 2002. "Identification and Identity." In S. Buss & L. Overton (Eds.), *The Contours of Agency: Essays on Themes from Harry Frankfurt*, (pp. 99–123). Cambridge, MA: MIT Press.

Venkatesh, S. 2000. *American Project: The Rise and Fall of a Modern Ghetto*. Harvard, MA: Harvard University Press.

Venkatesh, S. 2008. *Gang Leader For A Day: A Rogue Sociologist Takes to the Streets*. New York: Penguin Books.

Vohs, K. 2006. "The psychological consequences of money." *Science*, 314: 1154–1156.

Vranas, P. 2005. "The Indeterminacy Paradox: Character Evaluations and Human Psychology." *Noûs*, 39: 1–42.

Waldron, J. 2014. "It's All For Your Own Good." *New York Review of Books*, October 9.

Walinsky, A. 1987. "What It's Like To Be In Hell." *New York Times*, Opinion: December 4, 1987.

Walker, O. 1969. "Why Should Irresponsible Offenders Be Excused?" *Journal of Philosophy*, 66: 279–290.

Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Wallace, R. J. 2007. "Moral Psychology." In F. Jackson and M. Smith (Eds.), *Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Wallace, R. J. 2014. "Reasons, Policies, and the Real Self: Bratman on Identification." In M. Vargas and G. Yaffe (Eds.), *Rational and Social Agency: The Philosophy of Michael Bratman*, (pp. 106–128). Oxford: Oxford University Press.

Walther, E. 2002. "Guilty by Mere Association: Evaluative Conditioning and The Spreading Attitude Effect." *Journal of Personality and Social Psychology*, 82: 919–34.

Walther, E., & Trasselli, C. 2003. "I like her, because I like myself: self-evaluation as a source of interpersonal attitudes." *Experimental Psychology*, 50: 239–246.

Washington, N., and Kelly, D. 2016. "Who's Responsible for This? Moral Responsibility, Externalism, and Knowledge about Implicit Bias." In M. Brownstein and J. Saul (Eds.), *Implicit Bias & Philosophy, Vol. 2*, (pp. 11–36). Oxford: Oxford University Press.

Watson, G. 1975. "Free Agency." *Journal of Philosophy*, 72: 205–220.

Watson, G. 1996. "Two Faces of Responsibility." *Philosophical Topics*, 24: 227–248.

Watson, G. 2001. "Reasons and responsibility: Review essay on John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*." *Ethics*, 111: 374–394.

Watson, G. 2002. "Volitional Necessities." In S. Buss & L. Overton (Eds.), *The Contours of Agency: Essays on Themes from Harry Frankfurt*, (pp. 129–159). Cambridge, MA: MIT Press.

Webb, T., Sheeran, P., and Pepper, J. 2010. "Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials." *British Journal of Social Psychology*, 51: 13–32.

Wedell, D. 1991. "Distinguishing among models of contextually induced preference reversals." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17: 767–778.

Wegner, D. 1989. *White bears and other unwanted thoughts*. New York: Viking.

Wegner, D. 2002. *The illusion of conscious will.* Cambridge, MA: MIT Press.

Wegner, D. 2008. "Self is magic." In J. Baer, J. Kaufmann & R. Baumeister (Eds.), *Are we free? Psychology and free will*, (pp. 226–247). New York: Oxford University Press.

West, P. 1996. "Predicting preferences: An examination of agent learning." *Journal of Consumer Research*, 23: 68–80.

Wheatley, T., and Haidt, J. 2005. "Hypnotic disgust makes moral judgments more severe." *Psychological Science*, 16: 780–784.

White, M. 2013. *The manipulation of choice: Ethics and libertarian paternalism*. New York: Palgrave Macmillan.

Williams, B. 1981. "Persons, Character and Morality." In *Moral Luck*, (pp. 1–19). Cambridge: Cambridge University Press.

Williams, L., and Bargh, J. 2008. "Experiencing Physical Warmth Promotes Interpersonal Warmth." *Science*, 322: 606–7.

Wilson, T. 2002. *Strangers to ourselves: Discovering the adaptive unconscious.* Cambridge, MA: Belknap Press.

Wilson, T., Lindsey, S., and Schooler, T. 2000. "A model of dual attitudes." *Psychological Review*, 107: 101–26.

Wilson, W. J. 1980/2012. *The Declining Significance of Race: Blacks and Changing American Institutions*, 3rd ed. Chicago: University of Chicago Press.

Wilson, W. J. 1987/2012. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Pol-*

*icy*, 2nd ed. Chicago: University of Chicago Press.

Wilson, W. J. 1996. *When Work Disappears: The World of the New Urban Poor*. New York: Knopf.

Wilson, W. J. 2009. *More Than Just Race: Being Black and Poor in the Inner City*. New York: Norton.

Wolf, S. 1987. "Sanity and the Metaphysics of Responsibility." In F. Schoeman (ed.), *Responsibility, Character and the Emotions*, (pp. 46–62). Cambridge: Cambridge University Press.

Wolf, S. 1990. *Freedom within Reason*. Oxford: Oxford University Press.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J., & Hitchcock, C. 2003. "Explanatory Generalizations, Part I: A Counterfactual Account." *Noûs*, 37: 1–24.

Zhang, H., & Chan, D. 2009. "Self-esteem as a source of evaluative conditioning." *European Journal of Social Psychology*, 39: 1065–1074.

Zheng, R. 2016. "Attributability, Accountability, and Implicit Bias." In M. Brownstein and J. Saul (Eds.), *Implicit Bias & Philosophy, Vol. 2*, (pp. 62–89). Oxford: Oxford University Press.

Zimbardo, P. 2007. *The Lucifer effect*. New York: Random House.

Zorbaugh, H. 1929. *The Gold Coast and the Slum*. Chicago: University of Chicago Press.