CHAPTER 13

AUTONOMOUS AGENCY AND SOCIAL PSYCHOLOGY

Eddy Nahmias

Autonomous agents, like autonomous nations, are able to govern themselves. They are not controlled by external forces or manipulated by outside agents. They set goals for themselves, establishing principles for their choices and actions, and they are able to act in accord with those principles. Just as deliberative democracies legislate so as to balance competing interests, autonomous agents deliberate to reach some consistency among their competing desires and values. And just as good governments create their laws in the open without undue influence by covert factions, autonomous agents form their principles for action through conscious deliberation without undue influence by unconscious forces. Autonomous agents are self-controlled not weak-willed, self-aware not self-deceptive. Given this description, it would be nice to be an autonomous agent.

Indeed, we believe we *are*, for the most part, autonomous agents.[1] However, there are threats to this commonsense belief. Some philosophers argue that if causal determinism is true then we lack free will and hence are not fully autonomous or responsible for our actions.[2] One might also worry that if certain explanations of the mind-body relationship are true, then our conscious deliberations are epiphenomenal in such a way that we are not really autonomous. Philosophers also analyze political freedom and various socio-political threats to people's autonomy.

But other threats to autonomy are less often discussed, threats that are not metaphysical or political but psychological. These are threats based on putative facts about human psychology that suggest we do not govern our behavior according to principles we have consciously chosen. For instance, if our behavior were governed primarily by unconscious Freudian desires rather than by our reflectively considered desires, we would be much less autonomous than we presume. Or if our behaviors were the result of a history of Skinnerian reinforcement rather than conscious consideration, our actions would be shaped by our environment more than by our principles. Since the influence of Freud and Skinner has waned, we might feel we have escaped such threats to our autonomy from the psychology. But, as I will explain below, more recent and viable theories and evidence from social psychology pose significant threats to autonomous agency.

1. AUTONOMOUS AGENCY

In this section I explain more fully what I mean by autonomous agency. In the following section, I outline the relevant research in social psychology, the ways it threatens our autonomy, and some responses to these threats.

On most conceptions of autonomous agency it clearly requires freedom of action, which is the ability to act on one's desires without external constraint. But such freedom is not sufficient for autonomy, since agents may also be *internally* constrained or influenced by external factors in ways more subtle than constraint or coercion. For instance, agents may act on addictions, phobias, or even strong passions that they would prefer not to move them. Or agents may be influenced by subtle manipulations (e.g., advertising) or by unrecognized situational factors (e.g., peer pressure) that, had the agents known about them, they would not want to influence them. These cases suggest that autonomy requires more than simply being free to act on one's desires; it also requires some measure of internal consistency among one's desires and values and some capacity to understand oneself and one's situation. I will discuss these requirements in terms of *principles* and *knowledge*.[3]

1.1 Autonomy requires reflectively chosen principles

We adult humans are autonomous in ways that young children and animals are not. This is in part because we are often moved not simply by our strongest urge in a given situation; we are also able to consider our immediate desires in terms of our long-term goals, including our moral and social obligations.[4] However, it would be inefficient, if not impossible, to reflect on our goals and obligations every time we act. Rather, we tend to deliberate about such matters calmly and reflectively to establish the ways we hope to respond *without* much reflection when faced with the relevant situations.[5] This deliberation may take the form of counterfactual reasoning: when I am in situations of type *X*, I should do *Y*. For instance, if I find myself confronted with a person in need, I should respond by helping him; as I consider job applicants I should ignore irrelevant information and focus on the information I deem important. The details may be left somewhat vague, but the goal is to establish *principles* for action, reasons that will guide you to act in particular ways in certain types of situations so that you act consistently with your reflectively considered preferences, even when you do not or cannot consciously deliberate at the moment of action.

Hence, autonomous agency requires the ability to form and act on principles. The formation of these principles should occur through conscious deliberation without the influence of any unconscious motivations the agent would reject if she knew about them. And the principles should be as internally consistent as possible so that the agent does not betray some of her own principles by acting on others.

Given this conception of autonomy, we can see that an agent's autonomy is threatened to the extent that she is *ignorant* of factors that lead her to act against her principles—i.e., were she to recognize these factors, she would reject them.

Similarly, an agent's autonomy is threatened by *rationalization*: cases in which the agent finds herself acting against her consciously chosen principles but retrospectively comes up with reasons to justify her action. Here, actions inconsistent with prior principles are explained (to herself and others) in terms of *post hoc* principles adjusted to fit the actions.[6] Of course, we are all subject to cases of ignorance and rationalization. But are we subject to these challenges to our autonomy to a greater extent than we realize?

1.2 Autonomy requires knowledge

These considerations further suggest that autonomy depends on the agent's ability to know her principles and to know how to act on them. Ideally, an autonomous agent can articulate to herself and others her principles for action. At a minimum, she can recognize whether she is acting on reasons she would accept were she to consider them. An agent's knowledge of her principles is unhelpful if she is unaware of the motivational states or external factors that lead her to act against them. Ignorance of such factors hinders the agent's ability to counteract their influence in order to act consistently with her principles. Finally, the agent has to know how to get herself to act according to her principles, even when she feels more motivated to act against them. Hence, autonomy requires some capacity to introspect accurately on one's motivations for action and to know why one acts as one does. To the extent that we do not know ourselves and our situations, we lack autonomy.

1.3 Autonomy comes in degrees

Notice that the above conditions are not meant to provide an *analysis* of autonomy; I do not know what the complete set of necessary and sufficient conditions for autonomy would be. Rather, I have described these conditions in terms of *degrees* of satisfaction. Intuitively, different agents seem to *possess* more or less autonomy, and an agent seems to *exercise* more or less autonomy in different decisions or actions. The degree to which agents are autonomous seems to align nicely with the degree to which they know their own principles for action and know how to act on those principles (though there are likely other conditions I have neglected). Accordingly, autonomy is compromised to the degree these conditions are compromised.[7]

Seeing autonomy in this way accords with our practices of attributing moral responsibility (e.g., praise and blame): we attribute responsibility to varying degrees depending in part on (1) the degree to which agents possess the relevant capacities to form and act on principles and (2) the degree to which agents have the opportunity to act on their principles. For instance, we hold children responsible to the degree that they have matured to understand their reasons for acting, and we hold adults responsible to the degree that they are in a position to know their obligations and to control their actions accordingly.[8]

Finally, this view allows us to see that empirical challenges to autonomy come in varying degrees. For instance, information about human psychology may suggest that we possess less autonomy than we think, without thereby suggesting we are *entirely* subject to forces beyond our control. Hence, one way to read the rest of my discussion is this: To the degree that social psychology's theories and experimental results suggest limitations to the capacities required for autonomy, to *that* degree our autonomy is compromised. My main goal is to bring attention to these largely unnoticed empirical threats to autonomy and to examine their depth and scope. These points will reinforce an underlying theme of this chapter: that autonomy should be investigated empirically as well as conceptually and that exploring empirical challenges to autonomy and responsibility is at least as illuminating as debating would-be global threats such as determinism.[9]

2. THE THREAT OF SOCIAL PSYCHOLOGY

We have seen, then, that an agent lacks autonomy to the extent that she is unable to know her own motivations or reasons for action or to know what situational factors are influencing her to act against her principles. Furthermore, an agent lacks autonomy to the extent that she acts in conflict with principles she has adopted or acts on reasons she would reject if she were to consider them.

Research in social psychology over the past few decades suggests significant limitations to these conditions of autonomy.[10] Specifically, some social psychologists have interpreted their research as demonstrating three interrelated theses:

(1) *The Principle of Situationism*: Our behavior is influenced to a significant and surprising extent by external situational factors that we do not recognize and over which we have little control. These factors are often ones we would not want to have such influence on us if we knew about them.

(2) *The Disappearance of the Character Traits*: Internal dispositional states are not robust or stable across various situations; traditional character traits are not good predictors of behavior. Hence, consistent principles we endorse or aspire to develop tend to be ineffective given the power of certain situational factors.

(3) *The Errors of Folk Psychology and Introspection*: We generally do not know about the first two theses, and hence our explanations of our own and others' behaviors are based on mistaken folk theories or inaccurate introspection. Our introspection does not give us privileged access to what motivates us to act.

2.1 Experiments and Implications

In order to clarify these theses, I will summarize some experimental results.[11] The most common experimental paradigm is simple. The psychologists manipulate certain factors that we would not expect to influence our behavior. But the

experimental group, which is exposed to the manipulated factor, behaves significantly differently than the control group, indicating the influence of that factor on behavior. Meanwhile, behavior *within* each group is consistent enough to suggest that other factors—including personality traits—play no significant role in determining behavior. Often, subjects are then asked to explain why they behaved as they did. They do not mention the manipulated factor as having played any role; rather, they mention other, more "principled" reasons for their choices and actions. Some of these experiments involve relatively trivial behavior, in which case subjects may feel obliged to come up with rationalizations for behaviors that are not generally guided by reflectively chosen principles.[12] However, other experiments involve situations in which most people presumably seek to act in accord with their reflectively considered principles. I will focus on experiments involving morally relevant behavior.

1) In 1964 when Kitty Genovese screamed for help for half an hour while being stabbed and raped in Queens, the forty people who witnessed the event did not help or even call the police. Social psychologists began testing whether this lack of intervention may have been due to a situational factor—the number of bystanders present—rather than, as the media explained it, the inherent apathy and callousness of New Yorkers. Numerous experiments showed that increasing the number of people who witness an emergency or a person in distress decreases the chances that anyone will intervene. For instance, when subjects heard a woman take a bad fall, 70% of solitary subjects went to help, but if subjects sat next to an impassive confederate, only 7% intervened.[13] A plausible explanation is that when we are around others, our perception of the situation alters; perceived responsibility to act is diffused by the possibility that someone else will (or might) take action.[14] Confounding the problem, if no one does take action, we construe the situation as less serious—if no one is reacting, it must not be so bad after all.[15]

But people do not *recognize* these group effects: "We asked this question every way we knew how: subtly, directly, tactfully, bluntly. Always we got the same answer. Subjects persistently claimed that their behavior was not influenced by the other people present. This denial occurred in the face of evidence showing that the presence of others did inhibit helping".[16] Rather, when asked, subjects, like the media, refer to dispositional traits (e.g., apathy or altruism) to explain their own and others' behavior—traits that do *not* significantly correspond with people's behavior—or they refer to their perception of the situation, which is skewed by a situational factor they do not recognize as influencing them. Presumably, people's refusal to admit the influence of group effects is an indication that they do not accept it as a legitimate influence. Rather, the principles they articulate refer to dispositions to respond to those in need based primarily on how great the perceived need is. Hence, it appears that people can fail to act on their principles because of situational factors they don't recognize or accept as reasons to act.

2) In another experiment Princeton seminary students were asked to prepare a lecture either on the parable of the Good Samaritan or on their job prospects. Some

subjects were told they were late getting to the lecture hall while others were not. En route, they came upon a man slumped in a doorway, coughing and groaning (as in the Biblical story). While 63% of the "early" subjects offered help, only 10% of the "late" subjects assisted the man in need. No significant correlations were found between the subjects' helping behavior and their self-reported personality traits or the subject matter of their lecture.[17] The "hurry" factor influenced some subjects by changing their perception of the situation: "because of time pressures, they did not perceive the scene in the alley as an occasion for ethical decision".[18] Again, people are not aware of the influence of this situational factor on their perception or their behavior.[19] Even if people consider themselves altruistic, even if they prefer *not* to be affected by factors (like being in a hurry) that they view as irrelevant to helping those in need, it is difficult to see how they can consciously override the influence of factors they do not believe influence them.

3) A study by Isen and Levin showed that subjects who found a dime in a payphone, and hence got a "mood boost", were then fourteen times more likely to help a passerby pick up dropped papers than subjects who did not find a dime.[20] Again, no one predicts or desires that their helping behavior is influenced by such seemingly irrelevant factors.

4) Finally, the well-known Stanley Milgram obedience studies consistently found that about two-thirds of subjects will shock a man into unconsciousness during a learning experiment.[21] Situationists suggest that the incremental levels of the shocks (15-volts) make it difficult for subjects to find a justifiable point at which to question the authority of the experimenter.[22] People certainly do not predict of themselves that they would continue well past the point that the learner appears to go unconscious to the 450-volt switch marked "Danger: XXX". We assume our principles would preclude us from performing such actions. And no one predicts that so many others would do it either.[23]

In each of the above experiments people's explanations for their own and others' behavior refer to character traits or principled reasons while ignoring the situational factors that in fact make the significant difference. That is, people are *ignorant* of significant causes of their behavior and, if asked to explain their behavior, they tend to offer *rationalizations*, confabulating reasons to try to make sense of their behavior.[24] Psychologist Roger Schank summarizes the general idea: "We do not know how we decide things […]. Decisions are made for us by our unconscious; the conscious is in charge of making up reasons for those decisions which sound rational".[25]

To the extent that situational factors play a large role in determining our behavior, differences in people's internal dispositions appear to play a relatively small role. That is, if a particular situational factor (such as finding a dime) can elicit similar behavior from most subjects, then, it is claimed, differences between individuals' characters are correspondingly insignificant in producing their behavior. The question of whether this work in social psychology implies an *elimination* of

character traits is controversial both within social psychology and in philosophical reactions to it.[26] I will not try to adjudicate that debate. I will simply suggest that even if *elimination* is not warranted, *to the extent* that character traits play less of a role in our behavior that we ordinarily suppose, this would threaten our autonomy to the extent we do not then act on consistent principles we endorse.[27]

In addition to experiments like those described above, the evidence for the "disappearance of character traits" comes from experiments that test for correlations in subjects' behavior across various situations designed to elicit trait-relevant responses (e.g., honesty, extroversion, impulsivity). For instance, Hartshorne and May tested students for honesty by examining their willingness to steal money, lie, and cheat on a test.[28] While the subjects behaved similarly in repeated cases of any *one* of these situations, they did not behave consistently *across* these situations; for instance, knowing that a particular student cheated on a test did not reliably indicate whether he would steal or lie. In general, people predict a very high correlation between character trait descriptions and behaviors in relevant "trait-eliciting" situations, but such correlations in fact hit a very low "predictability ceiling".[29] In other words, if we want to understand why an agent does what he does in situation *X*, we are better off either looking at his past behavior in situations just like *X* or at the way most people behave in *X* than we are considering what we take to be the agent's relevant character traits.

Such conclusions raise a problem for autonomy because the principles we adopt for our own behavior are usually not so situation-specific; rather they look more like character traits that will dispose us to feel and respond to a wide range of situations in an appropriate way. To act on a principle of honesty requires overcoming inclinations to lie or cheat across a range of situations where such behavior is inappropriate. We do not usually form situation-specific principles, such as the desire to help people specifically when we are not in a hurry.[30] Rather, we aim to develop consistent tendencies to respond to the aspects of situations that call for traits such as altruism, generosity, courage, or diligence. The more these tendencies can be disrupted by unrecognized variations in situational factors that we don't want to influence us, the less consistent we will be. And the less we recognize how easily these tendencies can be disrupted, the less we can make conscious efforts to shore them up in order to act consistently.[31]

To the extent that the dispositional traits we identify as our principles do not in fact correspond with consistent behavior, we are identifying ourselves with constructed concepts rather than actual motivational states. This limits the influence of our principles on our actions. If future behavior cannot be accurately predicted based on the possession of a character trait, then adopting a principle to act on that trait appears fruitless. I may reflectively endorse being generous but doing so appears to be ineffective to the extent that (1) I cannot predict how I will be influenced by situational factors that vary across different "giving" opportunities, and (2) there is no such trait as generosity to cultivate. Research in social psychology suggesting the disappearance of character traits thus threatens our autonomy to the extent that it suggests limitations on our capacities to form consistent principles for action and to know how to act on our principles.

Such research also suggests a more systematic limitation to our knowledge. If we tend to explain human behavior in terms of character traits while being ignorant of the influence of unnoticed situational influences, then this suggests that our explanations and predictions of our own and others' behavior will often be inaccurate. Furthermore, on this view, our introspective reports about why we feel and act as we do are based on our (largely inaccurate) folk theories rather than any direct access to our own mental processes. Social psychologists argue that our folk psychology suffers from the *fundamental attribution error*, "people's inflated belief in the importance of personality traits and dispositions, together with their failure to recognize the importance of situational factors in affecting behavior".[32] Then, they suggest, when we introspect on why we feel or act as we do, we grasp for the most plausible folk psychological explanation, rather than accurately introspecting on our own mental states: "The accuracy of subjective reports is so poor as to suggest that any [introspective] access that may exist is not sufficient to produce generally correct and reliable reports".[33]

There are numerous social psychology experiments suggesting that our folk psychology and our introspection are inaccurate.[34] I'll describe just one to illustrate the threat to autonomy. Nisbett and Bellows asked college students to assess a candidate, Jill, for a counseling job.[35] Subjects judged Jill according to four criteria ("likeability", "intelligence", "sympathy", and "flexibility") after they read a three-page application file with information about her life, her qualifications, and a prior interview with her. Five factors in Jill's file were manipulated across different sets of subjects: (1) whether Jill was described as attractive, (2) whether she had superior academic credentials, (3) whether she had spilled coffee during her interview, (4) whether she had recently been in a car accident, and (5) whether the subject was told he or she would later meet Jill. After judging Jill on the four criteria, subjects were then asked to introspect about how much each of these factors had influenced their judgments on each of the criteria.[36]

The results show that the subjects' introspective reports rarely correlated with the actual effect the various factors had, as determined by between-subject comparisons (with one exception: academic credentials did influence judgments of intelligence just as subjects reported). (See Figure 13.1). For instance, subjects reported that whether Jill had been in an accident had the greatest effect on their judgments of her sympathy and that her academic credentials had the greatest effect on how much they liked her, but when the manipulated factors were compared across subjects' judgments, it turned out that these introspective reports were inaccurate; instead, believing they would meet Jill had, by far, the strongest effect on subjects' judgments of her sympathy (as well as her flexibility) and reading that Jill spilled coffee in her interview had the strongest effect on how much they liked her!

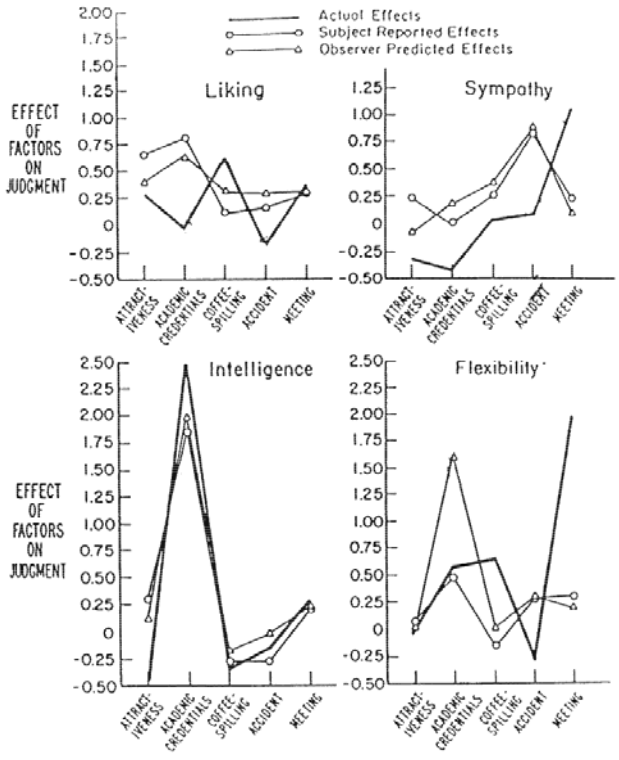

Figure 13.1. Nisbett and Bellows' (1977) experiment. Bold line represents actual effects of manipulated factors (on X-axis) on judgments (Y-axis) as measured across "actor" subjects; circle line represents reports by "actor" subjects about the effects of the factors on their judgments; and triangle line represents predictions by "observer" subjects about the effects the factors *would* have on their judgments. Reproduced from Nisbett and Bellows (1977, 619).

A second group of subjects was asked simply to imagine they were judging some unspecified candidate for an unspecified job and, without actually viewing *any* information, they rated how important the five factors (appearance, academics, etc.) would be to their judgments on the four criteria (likeability, sympathy, etc.). These subjects' predictions were statistically identical to—and hence as inaccurate as—the reports by the subjects who actually considered the information about Jill (see Figure 13.1). The authors interpret these results as showing that the introspecting subjects did not actually make their judgments based on the reasons they reported. Rather, they were unaware of how various factors influenced most of their judgments, and they retrospectively theorized about what influenced them in the same way the uninformed subjects theorized about what factors *would* influence them. And in both cases, the theories were generally mistaken. In the one case where

the theories were accurate, it is because the connection between academic credentials and intelligence fits plausible cultural norms about such judgments: the authors explain, "[v]erbal reports should be correct wherever there exists a correct causal theory, correctly applied in the particular instance",[37] but "introspection played little part in subject reports".[38] Social psychologists hence suggest that our introspection about why we do what we do looks more like *theoretical* reasoning about what someone might do in our circumstances.

Nisbett and Wilson conclude from such experiments: "It is frightening to believe that one has no more certain knowledge of the workings of one's own mind than would an outsider with intimate knowledge of one's history and of the stimuli present at the time the cognitive processes occurred".[39] Indeed it *is* frightening, because our autonomy is compromised if it turns out that what we think we are doing when we introspect on the reasons we act does not in fact involve reliable access to our principles, but instead involves "just so" stories derived from our inaccurate folk psychology. The threat appears graver still if psychologists develop theories to predict our own behavior *better* than we ourselves can, since prediction goes hand in hand with control.

Let me be more precise about the threats to autonomy posed by such research. In these experiments, when subjects offer reasons for their attitudes and actions, they are usually doing two things: (1) claiming that those factors have causally influenced them, *and* (2) explaining the factors they think *justify* their attitudes and actions. For example, when subjects report that knowing Jill has been in an accident made them more likely to see her as sympathetic, they are presumably explaining not only the influence of that factor but also their view that its influence is legitimate (e.g., she'll understand suffering better). Conversely, when subjects report that Jill's spilling coffee was *not* a factor in their rating of how much they liked her, they are reporting that they think it had no influence on them and also that they think it *should* have no influence on them—clumsiness is not a good reason to like (or dislike) Jill.[40] In such cases, the reasons subjects report as the basis of their judgments often accords with principles they accept (or would accept). However, these experiments indicate that factors the subjects see as irrelevant are in fact influential while most factors they see as important are *not*. To the extent such results can be generalized, it seems that the reasons we offer as explanations for our behavior look more like retrospective rationalizations.

Furthermore, such experiments suggest limitations on our knowledge of how to influence our actions to accord with our principles. For each unrecognized effect on our motivations and actions, our deliberations about how to influence our future actions are correspondingly restricted. For instance, if you're hiring someone for a job, you'd likely deliberate about which criteria are important to you and how you will determine if candidates meet those criteria. But it seems your judgments may often be affected by unrecognized factors that you would not want to be influential, such as the candidate's appearance or whether they spill coffee at the interview.[41] The less accurate your knowledge of which factors might influence you, the less you can control the influence of those factors you want to make a difference—that is, the less you can act on your principles.

2.2 Responses to the Threat of Social Psychology

I have presented these experimental results and interpretations from the perspective of the social psychologists in order to highlight the challenges they pose to autonomous agency. But I will now suggest some avenues for responding to their interpretations so that we might defend our autonomy against these threats.[42] My main goal has been to show why the degree to which we are autonomous agents is, in large part, an empirical question. But I will now suggest that it remains, in large part, an open question.

First of all, the extent of human actions to which the social psychology evidence applies remains unknown. For instance, these experiments usually involve complex experimental set-ups, designed precisely to "trick" subjects, who are doing things without being asked to attend to what they are doing. In some cases there is little reason to think subjects care about the activities being studied, so they may not really have what they see as reasons for what they do—they only come up with reasons when asked to.[43] Perhaps some of our more considered actions are less subject to situational effects. And some the situational effects will surely involve the salient aspects of the situations, the very ones we get ourselves into by consciously choosing to do so. For instance, when someone deliberatively settles on becoming an ER doctor, they are knowingly putting themselves in situations that will lead them to help people in distress. A reflective decision to become a professional philosopher includes a decision to be in many fewer such situations.

Furthermore, in the experiments subjects are asked to explain their judgments or responses only *after* they complete them. When the subjects report their experiences, they are not introspecting but *retro*specting on processes they performed earlier during the task. Perhaps subjects' poor memory of the thoughts they had accounts for some of the problems ascribed to their poor introspection.[44] None of these experiments ask subjects to consider what principles they want to influence them *before* they act, in order to determine whether such deliberation can counteract the disconnect between subjects' reported reasons and their actions. It would be helpful to test what would happen if subjects were asked to consider their principles of action before they engage in the relevant behavior.[45] Experiments that test for the effect of prior introspection are needed to determine the influence of conscious deliberation about one's principles.

In fact, some experiments *have* tried to determine the effect of one type of prior introspection on one's behavior. These experiments, however, suggest that introspecting on the reasons for our attitudes can be disadvantageous. Specifically, they imply that when subjects introspect about why they feel the way they do about something, they may not access the actual reasons for their feelings but instead come up with what they think are plausible reasons, and this disrupts the consistency between the feelings they then report and their subsequent actions, and they may even make choices they later regret. The upshot, according to Timothy Wilson, is that "self-reflection may not always be a beneficial activity"; indeed, "at least at times, the unexamined choice *is* worth making".[46]

For example, one experiment asked dating couples to report their feelings about their relationship, including how long they thought it would last. But one set of subjects was first asked to introspect on the reasons for their feelings about their relationship, while control subjects just reported their attitudes without introspecting on them. The correlation between subjects' reported attitudes (e.g., how long they thought the relationship would last) and their behavior (i.e., whether they were still dating several months later) was significantly lower (.10) for those subjects that introspected than for the controls (.62).[47] As Wilson interprets it, the introspecting subjects came up with what they saw as plausible reasons for their feelings but they did not have direct access to the actual reasons they felt as they did. These subjects then adjusted their reported attitudes to match the reasons they had thought up so that their behavior, motivated by their "real" attitudes, did not then match their introspective reports.

The problem raised by this experiment (and others like it[48]) is not only that it provides more evidence that we are often mistaken in our explanations for our actions and attitudes, but also that it suggests we often act on motivations we have not considered and which, in fact, we might not accept as principles for action. Luckily, in this case, there is actually some experimental evidence that limits the scope of such research. In the dating study and others like it, when the experimenters controlled for subjects' knowledge (e.g., whether the partners knew each other well), they found that knowledgeable subjects did *not* face the problem of inconsistency after introspecting on their attitudes. That is, the behavioral measure aligned much more closely with the attitudes subjects reported based on their introspected reasons.[49] Such studies suggest that when subjects know and care about the relevant issue, they seem to have *already* considered the principles they want to motivate their behavior. So, unlike unknowledgeable subjects, they need not adjust their attitudes to match reasons they come up with on the spot. They already have good reasons and their attitudes and behaviors reflect these reasons. In such cases, it seems we have deliberated about our reasons for feeling and acting as we do, and these deliberations have "sunk in" so that our actions align with our principles. Our introspection on our reasons does not disrupt our attitudes—rather, we revisit a pattern of reasoning we have already made our own.

This is consistent with my account of autonomy, because it suggests that increased knowledge of the world and ourselves increases our ability to act in accord with our principles.[50] And it suggests that reflective consideration of one's principles is the first step in getting oneself to act on them. When unknowledgeable subjects introspect on their reasons, they are trying to locate their reasons for the first time, trying to justify their attitudes; it is not surprising that their behavior does not immediately match these reasons. But such introspection may initiate the process of disrupting and overcoming habitual behavior and unconsidered motivations that don't accord with the principles the agent would accept.[51]

My reason for examining the social psychology research has not been to conclude that it shows we are not autonomous agents. Rather, my aim has been to bring attention to some implications of social psychology that have not been fully examined and to illustrate that the scope of our autonomy can and should be

examined empirically. While the research I have examined does suggest important limitations to our autonomy, more experiments are required to learn when and how conscious consideration of our principles makes a difference in our behavior. Too few experiments have dealt with attitudes and actions we care about, whose outcomes are of direct relevance to our significant interests and goals. And too few have examined how prior deliberation about how we want to be influenced affects what we in fact do.

But regardless of these shortcomings, social psychology offers useful paradigms and starting points for the empirical investigation of what has for too long been designated a merely conceptual issue, the nature and scope of our autonomy.[52]

NOTES

[1] Below I explain why autonomy is best understood as a property agents may possess and exercise to varying degrees.

[2] The connections between the concepts of autonomy, free will, and moral responsibility are complicated, in part by the variety of ways philosophers use each of them. Though I will not defend it here, I believe an adequate account of autonomous agency will be sufficient as an account of free and morally responsible agency, and my discussion may be read in this way. See Taylor (2005).

[3] For accounts of autonomy that suggest some of the features I outline, see Frankfurt (1988; 1999), Dworkin (1988), Taylor ([1977] 1982), Wolf (1990), Watson (1975), Mele (1995), Christman (1991), Bratman (1987) and Fischer and Ravizza (1998). Some of these accounts, however, are described in terms of conditions required for free agency or morally responsible agency rather than autonomous agency (see note 2).

[4] That we often take on such obligations without coercion suggests, somewhat paradoxically, that autonomy can include being governed by a form of external control so long as the agent autonomously accedes to such control.

[5] Regret facilitates such deliberation about how to act differently next time around. Watson (1975) discusses the importance of acting on one's *values*, "those principles and ends which [the agent]—in a cool and non-self-deceptive moment—articulates as definitive of the good, fulfilling, and defensible life" (105).

[6] Such rationalization may be difficult to distinguish from cases where the agent *modifies* her principles in light of her actions or their outcomes. One way to test whether an agent is rationalizing her actions with principles she does not really hold is to see whether or not she accepts those principles at other times and in relevantly similar situations.

[7] Cases become complicated when the agent knowingly chooses to do something that will compromise his opportunity to satisfy these conditions in his later actions. For instance, he may autonomously take drugs knowing it will compromise his ability to act autonomously. Hence, we sometimes attribute responsibility to an agent's actions that were not autonomously performed (e.g., when he lacks knowledge or control), because those actions are "traceable" to actions or choices that he did perform autonomously (e.g., many cases of drunk driving).

[8] This explains why we mitigate an agent's responsibility when he is ignorant of his obligations or the consequences of his actions (where such ignorance is not itself culpable—see previous note), or when he is under extreme emotional duress or cognitive load.

[9] Elsewhere I examine similar empirical challenges to autonomy (or free will) from other sciences that explore human nature, such as evolutionary psychology or neurobiology, each of which suggests certain limitations on our knowledge of and control over our motivations and behaviors. See, e.g., Libet (1983), Wegner (2002), and Nahmias (2002).

[10] I should note that social psychology is a diverse field; the work I discuss represents selected elements within it, notably the "situationist" camp led by researchers such as Lee Ross, Richard Nisbett, Timothy Wilson and their collaborators. For objections to the situationist paradigm, see, e.g., Sabini, Siepmann, and Stein (2001), Krueger and Funder (2004), and Cotton (1980). For more detailed discussions of the implications of this research for free will and responsibility, see Nahmias (in prep.), Doris (2002) and Nelkin (in press).

[11] For extensive reviews of such experiments, see Nisbett and Wilson (1977), Ross and Nisbett (1991), and Wilson (2002).

[12] For example, the position effect: presented with identical products, consumers tend to select the ones to the right, but they reject the role of position in their decision, instead coming up with reasons why they think the selected product is better than the other (identical) ones. See Ross and Nisbett (1991, 30-32). See Spinner (1981) for discussion of the distorting effects of demand conditions (i.e., subjects' being asked to explain their actions).

[13] See Latane and Darley (1968, 1970) and Ross and Nisbett (1991, 42). In another experiment, when *solitary* subjects heard an experimenter feign an epileptic seizure, 85% intervened; when subjects believed there was one other subject listening, 62% intervened; when they believed there were four other subjects, 31% intervened. And in all these experiments, interventions occurred faster when there were fewer subjects.

[14] A woman who heard Genovese screaming explained, "I didn't let" my husband call the police; "I told him there must have been 30 calls already".

[15] Post-experiment interviews suggest this interpretation: subjects in groups describe the emergencies in different terms (e.g., the fall victim's "cries" become "complaints") and notice them more slowly than subjects who are alone. Subjects also may want to avoid embarrassing themselves by taking action when no one else seems to think something should be done, though in some experiments subjects could not even see how others were reacting.

[16] Latane and Darley (1970, 124).

[17] Darley and Batson (1973).

[18] Ibid., 108.

[19] See Pietromonaco and Nisbett (1982); they describe to subjects the Good Samaritan study and then ask them to predict the outcomes. Subjects predict that the *majority* of seminary students would stop to help in all conditions, but that 20% more would help if their religious calling was based on a desire to help others. Subjects thought that being in a hurry would make *no* difference to whether the seminary students helped.

[20] Isen and Levin (1972). See Miller (2004, appendix) for some confounding factors regarding these experiments.

[21] See Milgram (1969).

[22] See Ross and Nisbett (1991, 56-58). This interpretation is strengthened by the fact that of the subjects who *do* stop the experiment, most do so when the "learner" appears to go unconscious (at 300 volts), a point where they can offer a justification for stopping. Many other factors have been shown *not* to correlate with subjects' behavior in the experiment, including gender, age, socioeconomic status, national origin, and various personality measures.

[23] After having the experiment described to them, *psychologists* predicted that only 2% of the subjects would continue to the end.

[24] However, in some cases subjects act so inconsistently with their principles that they disassociate themselves from their behavior, expressing surprise and dismay at their actions. This happened with some of Milgram's subjects, and in another famous situationist study, the Zimbardo prison experiments, where subjects given the role of prison guard became so cruel and aggressive that the experiment had to be terminated. Though some guards offered justifications for their behavior, one self-described pacifist said of his force-feeding a prisoner, "I don't believe it is me doing it," and another said, "I was surprised at myself. I was a real crumb" (Doris 2001, 51-53). Comparisons with the behavior of American soldiers at Abu Ghraib seem apt (see Nelkin, in press). In fact, Zimbardo recently testified in their defense and wrote that the prison guards "had surrendered their free will and personal responsibility during these episodes of mayhem […]. [They] were trapped in a unique situation in which the behavioral context came to dominate individual dispositions, values and morality to such an extent that they were transformed into mindless actors alienated from their normal sense of personal accountability" (www.edge.org).

[25] Quoted at www.edge.org.

[26] See, e.g., Ross and Nisbett (1991, chapter 4); Doris (1998 and 2001); Harman (1999); Flanagan (1991, chapter 13); Merritt (2000); Miller (2003); Kamtekar (2004).

[27] Despite some similarities between the situationists and the more radical behaviorist tradition—notably the shared claim that environmental conditions play a significant role in human behavior—the situationists do recognize the importance of internal cognitive states, namely people's perceptual and motivational construal of their situations. However, these dispositions to perceive situations in certain ways are relatively specific and do not support the attribution of recognized character traits.

[28] Hartshorne and May (1928).

[29] Ross and Nisbett describe this "maximum statistical correlation of .30 between measured individual differences on a given trait dimension and behavior in a novel situation that plausibly tests that dimension [as] an upper limit. For most novel behaviors in most domains, psychologists cannot come close to that" (1991, 3).

[30] This is not to say that we don't aim to have principles that are open-ended and flexible, but we aim for them to be responsive to factors whose influence we accept or would accept, not to ones whose influence we don't recognize and would not accept if we did recognize it.

[31] It is this challenge to character traits that leads some philosophers—notably, Doris (1998 and 2001) and Harman (1999)—to suggest that social psychology challenges virtue ethical theories since they appear to require the possibility of robust character traits.

[32] Ross and Nisbett (1991, 4). That we tend to refer to character traits in explaining behaviors—especially others' behaviors—is a claim informed by experimental research, not just intuition (See ibid., chapter 5). The basic idea is that we *attend* to other people—dynamic and interesting as they are—and not to their environments—seemingly static and boring as they are—so that we *attribute* causal power to agents (the word "agent" itself suggests this) rather than to the situational factors that influence agents.

[33] Nisbett and Wilson (1977, 233). Their explanation for our successful cases in predicting others' behavior is that we usually interact with people in similar situations over time. And our explanations of our own behavior are accurate "due to the incidentally correct employment of [folk] theories" (233).

[34] See Nisbett and Wilson (1977); Ross and Nisbett (1991); Wilson (2002).

[35] Nisbett and Bellows (1977).

[36] Since each subject was exposed to four of the five factors, each offered 16 self-reports (one for each of the four factors' affect on each of the four judgments about Jill—i.e., likeability, intelligence, sympathy, and flexibility).

[37] Nisbett and Bellows (1977, 614).

[38] Ibid., 623.

[39] Nisbett and Wilson (1977, 257).

[40] Reported causal influences will not *always* be offered as justifications; for instance, some people may recognize that attractiveness influences their judgments of "likeability" but not accept that influence as a good reason.

[41] Ross and Nisbett (1991, 136-138) discuss the "interview illusion": whereas subjects believe interviews will correlate with performance at a rate of 0.6, the actual correlation between judgments based on interviews and later job performance is usually below 0.1.

[42] One response I will not explore here is that this research is empirically or conceptually flawed. For some suggestions that it is, see references in note 10.

[43] Many studies deal with consumer choices, or puzzles, or quickly made choices. In fact, I have found only one experiment, a dating study discussed below, in which subjects are reasoning about something that will have any direct impact on their own lives.

[44] Ericsson and Simon's *Protocol Analysis* (1984) shows that concurrent introspection is much more accurate than retrospection.

[45] For instance, in the job interview study, presumably, many subjects had at some point thought about the influence academic credentials should have on judgments of intelligence, and perhaps—just as they report—those thoughts did play a role in those judgments.

[46] Wilson and Schooler (1991, 192). See also Wilson (2002).

[47] Wilson *et al.* (1989). One might suggest that we do not expect or care to be able to offer accurate reasons for our romantic feelings (that the idea of *justifying* one's love seems wrong). However, we do offer reasons (to ourselves, friends, and family) for why we are in the relationships we are in, and presumably, we want those reasons to bear some relation to reality, and we do *not* want the act of coming up with such reasons to disturb our actual feelings.

[48] For instance, a study on voting behavior showed that subjects who introspected on why they liked or disliked candidates before reporting their attitudes thereby disrupted their reported attitudes so that their behavior correlated with their reported attitudes significantly less than it did in controls who did not introspect: -0.43 vs. 0.46 (Wilson *et al.* 1989).

[49] In the dating study, for introspecting subjects who knew their partner well, behavior correlated with reported attitudes at 0.56, compared with -0.19 for subjects who had been dating a short time (Wilson *et al.* 1984). In the voting study cited in the previous note, for subjects who knew a good deal about politics, the correlation between reported attitudes and behavior was 0.53 (Wilson *et al.* 1989).

[50] In fact, knowledge of the social psychology evidence itself might increase our autonomy (1) by improving our ability to attend to situational influences we tend to overlook (see Beaman, Barnes, and McQuirk 1978; see Pietromonaco and Nisbett 1992), (2) by informing us about how to manipulate our environment to increase the likelihood that we act on our principles (e.g., "channeling factors" are subtle changes in environment that can significantly increase

our ability to carry out planned actions), and (3) by decreasing our confidence in the strength of our character traits so that we do not rely on them to guide us successfully through tempting situations (see Doris 1998).

[51] See Holt (1989; 1993).

[52] This chapter is drawn from ideas developed over a number of years and with feedback from many audiences and friends to whom I am grateful. I would like to thank, in particular, Owen Flanagan, Al Mele, Dana Nelkin, John Doris, and Manuel Vargas.