Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
Vol. LXXXV No. 3, November 2012

What Makes a Manipulated Agent Unfree?

CHANDRA SEKHAR SRIPADA
University of Michigan, Ann Arbor

Incompatibilists and compatibilists (mostly) agree that there is a strong intuition that a manipulated agent, i.e., an agent who is the victim of methods such as indoctrination or brainwashing, is unfree. They differ however on why exactly this intuition arises. Incompatibilists claim our intuitions in these cases are sensitive to the manipulated agent's lack of ultimate control over her actions, while many compatibilists argue that our intuitions respond to damage inflicted by manipulation on the agent's psychological and volitional capacities. Much hangs on this issue because manipulation-based arguments are among the most important for defending incompatibilist views of free will. In this paper, I investigate this issue from a experimental perspective, using a set of statistical methods well suited for identifying the features of hypothetical cases people's intuitions are responding to. Results strongly support the compatibilist view—subjects' tendency to judge that a manipulated agent is unfree was found to depend on their judgments that the agent suffers impairments to certain psychological/volitional capacities that compatibilists say are the basis for free will. I discuss the significance of these results for the use of manipulation cases in the philosophical debate about free will.

1. Introduction*

The Argument from Manipulation, one of the most powerful in the incompatibilist's arsenal, is based on our intuitions about a family of hypothetical cases called 'manipulation cases'. In a standard manipulation case, one agent wants to get another agent to perform some action. But Manipulator does not force Victim to act as she desires, as would happen in coercion. Rather, Manipulator uses other methods of influence such as conditioning or social engineering to instill in Victim a comprehensive suite of desires, values, and other mental states, which

* Earlier versions of parts of this paper were presented at the Moral Psychology Research Group (Princeton University, March 2010). Special thanks to Eddy Nahmias for inspiration and very helpful comments.

thereby lead Victim to choose to perform Manipulator's preferred action *of Victim's own volition*. (Manipulator's intervention is typically supposed to be *covert* so Victim is unaware of the influence.) Suppose now that Manipulator is successful and Victim performs the action Manipulator intended by means of the mental states that Manipulator covertly shaped. Does Victim act freely?

According to many philosophers (Kane, 1996, Pereboom, 2001), there is a strong intuition that an agent in a standard manipulation case is not free, or at the very least that the agent's freedom is significantly diminished. Call this the 'manipulation intuition'. The question I address in this paper is what feature(s) of cases involving manipulation is the manipulation intuition sensitive to? Broadly speaking, two answers have been given. According to incompatibilists, the manipulation intuition is responsive to the agent's lack of ultimate control over her actions. An agent in a manipulation case chooses actions on the basis of her own desires and values. However, these desires and values are themselves comprehensively shaped by forces external to the agent, and in this sense, the agent is not the *ultimate* source of her actions. Free will, incompatibilists argue, requires that an agent exert ultimate control over her actions, and the manipulated agent's failure to possess this kind of control is what causes us to intuitively judge that she is unfree (Kane, 1985, Kane, 1996). Of course, incompatibilists are quick to add that the failure to be the ultimate source of one's actions equally afflicts agents in a deterministic world. This is why intuitions about manipulation cases, coupled with the appropriate auxiliary premises, have become a mainstay in contemporary defenses of incompatibilism (and I discuss the details of how such arguments work below).

According to some compatibilists, however, intuitions about manipulation cases are sensitive to something else altogether. Compatibilists typically argue that possessing free will amounts to being able to exercise certain key cognitive, evaluational, and volitional capacities. Many compatibilists have argued that exposure to manipulation damages or impairs these capacities in some way, thereby rendering the agent unfree. These compatibilists have proposed a number of different pathways by which manipulation can compromise the agent's psychological capacities (see Watson, 1987, Kapitan, 2000, Haji, 1998, Mele, 2006, Wolf, 1990, Dworkin, 1976, Greenspan, 1978, Shoeman, 1978, Scanlon, 1986, Shoemaker, 2003, Faraci and Shoemaker, 2010) and I discuss two specific pathways a bit later. But setting aside differences about the pathway(s) by which manipulation exerts these damaging effects, the key point is that many compatibilists share the view that intuitions in manipulation cases are not driven by the absence of ultimate control

(as incompatibilists claim), but rather these intuitions are sensitive to deficits in the agent's evaluative and volitional capacities brought about by manipulation.[1]

It is worth emphasizing again that the disagreement described above between incompatibilists and compatibilists over manipulation cases is, for the most part, *not* over the following question:

> (M1) Intuitively, is a manipulated agent, as described in a standard manipulation case, unfree?

This is a question about the *content* of our intuitions, and on this issue incompatibilists and most compatibilists are in agreement that the answer is 'yes', a victim of manipulation is indeed unfree. Where incompatibilists and compatibilists diverge is over a quite different question:

> (M2) What features of manipulation cases are our intuitions sensitive to?

Incompatibilists contend that our intuitions are responsive to the agent's failure to be the ultimate source of her actions, while compatibilists argue that our intuitions are responsive to impairments in certain critical psychological capacities. One might wonder, however, why (M2) has led to such heated debate and proven so intractable. After all, (M2) poses a quite specific question, and, thankfully, in contrast to many other philosophical questions, there is little disagreement among the disputants about what is the meaning of this question, or what would count as an appropriate answer. Why then has the debate about (M2) raged on for decades without any signs of resolution?

The answer, I believe, is that the task of addressing (M2) is an instance of a more general problem in philosophical inquiry that I have previously dubbed the 'the critical features problem' (Sripada and Konrath, 2011). This is the problem of figuring out which among the many features associated with a hypothetical case are the critical one(s) our intuitions are responding to. While armchair methods are quite good at telling us about the content of our intuitions (i.e., they are quite good at answering questions like (M1)), they are much less

[1] In section 1 below, I discuss a key corollary to the compatibilist position: To the extent that a manipulated agent is *not* viewed as having suffered damage to key psychological capacities, the agent is correspondingly *not* judged to be unfree.

effective for addressing question like (M2) that concern what features of hypothetical cases are driving our intuitions.

There are basically two reasons why questions like (M2) are difficult to answer from the armchair. First, in response to textual descriptions of hypothetical cases, people may construct very different mentally represented scenarios. This problem is readily apparent in the philosophical debate about manipulation cases in which compatibilists seem to be envisioning these cases as involving damage of one sort or another to the agent's psychological capacities while incompatibilists seem to be envisioning these cases as lacking these features. This kind of variability across individuals in how they envision hypothetical cases hampers efforts to get a clear sense of what feature(s) of these cases our intuitions are sensitive to. Second, even when individuals all imagine a hypothetical case in the very same way, because any philosophically interesting case inevitably involves the simultaneous presence of a large number of features, there is still the key question of which among the myriad features present in the case are the one(s) that are in fact driving our intuitions. This second limitation in the use of armchair methods for addressing the critical features problem is indeed very well supported by substantial quantities of psychological research, which shows that when people are placed in situations that produce strong intuitive reactions, they are quite unreliable at identifying which situational feature is the one driving their intuitive reactions (this extensive body of work is famously reviewed in Nisbett and Wilson (1977) and its relevance to philosophical theorizing is clarified and defended in Sripada and Konrath (2011)). In sum then, it is these two limitations in armchair methods for solving the critical features problem that, I believe, explain why questions like (M2) have proven so intractable and have generated so much heated philosophical debate.

In this paper, I address (M2) from an experimental perspective. In particular, I report the results of studies of ordinary people's intuitions in manipulation cases using certain statistical methods (described in detail below) that are especially well-suited to helping to uncover what features of hypothetical cases people's intuitions are responding to. This paper is divided into three parts. In part 1, I set up the Argument from Manipulation in greater detail, and formulate a 'Compatibilist Position' regarding what drives intuitions in manipulation cases. In part 2, I report the results of two studies of the intuitions of ordinary people in response to manipulation cases that are designed to test this Compatibilist Position. In part 3, I discuss the implications of the findings from these studies for the debate over manipulation between incompatibilists and compatibilists.

1. The Argument from Manipulation and Two Compatibilist Responses

1.1. The Argument from Manipulation

Manipulation arguments are aimed at a family of compatibilist views that distinguish agents who are free from agents who are not free by the presence of certain psychological conditions (hereafter called *compatibilist freedom-conferring psychological (CFP) conditions*). Individual compatibilist accounts differ somewhat on which are the psychological conditions that matter for an agent to be free. Some compatibilists emphasize the structure of an agent's motivations, for example, whether the agent's first-order desires conform to her higher-order volitions (Frankfurt, 1971), or reflect the verdicts of her valuation system (Watson, 1975). Other compatibilists focus on the presence of certain psychological abilities, such as the ability to will (Gert and Duggan, 1979) or the ability to respond to reasons (Fischer and Ravizza, 1998). The crucial insight behind manipulation arguments is that regardless of which psychological conditions compatibilists say are sufficient for an agent to be free, it seems at least possible that a manipulated agent could *fully* satisfy these conditions.[2] Consider the following case from Derk Pereboom in which Professor Plum kills Mrs. White due to manipulation by neuroscientists, and notice Pereboom's description of the manipulated agent's psychology:

> Plum is not constrained to act in the sense that he does not act because of an irresistible desire—the neuroscientists do not provide him with an irresistible desire—and he does not think and act contrary to character since he is often manipulated to be rationally egoistic. His effective first-order desire to kill Ms. White conforms to his second-order desires. Plum's reasoning process exemplifies the various components of moderate reasons-responsiveness. He is receptive to the relevant patterns of reasons, and his reasoning process would have resulted in different choices in some situations in which the egoistic reasons were otherwise. At the same time, he is not exclusively

[2] Some compatibilists seek to build extra historical conditions on top of standard CFP conditions specifically to block the ability of manipulated agents to count as being free (Fischer and Ravizza, 1998, Mele, 2006). It is unclear whether these approaches are entirely successful (e.g., see Kapitan, 2000 for a critique). In any case, as we shall see, the results of the studies reported in this paper suggest that these extra historical conditions are actually *unnecessary*. Standard CFP conditions are in fact sufficient to fully capture intuitions about free will in manipulation cases.

> rationally egoistic since he will typically regulate his behavior by moral reasons when the egoistic reasons are relatively weak—weaker than they are in the current situation. (Pereboom, 2001, pg. 112–113).

In this vignette, the manipulated agent exhibits a strikingly comprehensive repertoire of psychological attributes and capacities that compatibilists say are sufficient for free will. But because the psychological attributes and capacities by which the agent satisfies CFP conditions are caused to arise by means of external manipulation, the agent, it would intuitively seem, is *not* free. It follows then, according to this line of reasoning, that fulfilling CFP conditions is not sufficient for free will.

To sum up, manipulation arguments have the following general structure:

> **The Argument from Manipulation**
>
> Premise 1: S fulfills all the CFP conditions that compatibilists say are sufficient for being free.
>
> Premise 2: S's CFP-relevant psychological states were caused to arise by manipulation.
>
> Premise 3: S is not free.

Conclusion: Fulfilling CFP conditions is not in fact sufficient for being free.[3]

---

[3] This version of the Argument from Manipulation closely resembles the formulations in Kapitan (2000) and McKenna (2008). There is another version of the argument (see for example Vihvelin, 2007), that goes, very roughly, as follows:

> **Argument from Manipulation (version 2)**
>
> Premise 1: A manipulated agent is not free.
>
> Premise 2: There is no relevant difference between a manipulated agent and an agent in a deterministic world.
>
> Conclusion: An agent in a deterministic world is not free.

Here, the second premise is contentious. This premise will be true only if a manipulated agent fulfills all CFP conditions. If a manipulated agent fails to fulfill CFP conditions, as many compatibilists contend, then premise 2 fails. But the question of whether a manipulated agent fulfills CFP conditions is simply premise 1 of the first version of the Argument from Manipulation. I conclude version 1 of the Argument from Manipulation is more fundamental, and version 2 stands or fails only if version 1 does.

What do manipulation cases such as ones put forward by Derk Pereboom, Robert Kane, and Eleonore Stump[4] have to do with the Argument from Manipulation? The most plausible answer is that intuitions about manipulation cases are the basis for premise 3. That is, when presented with a hypothetical case involving a manipulated agent, the fact that we intuitively judge the agent to be unfree provides good (albeit defeasible) evidence that the agent *is* unfree. Now there may be other strategies by which premise 3 could be defended that do not rely on intuitions about manipulation cases. However, in what follows, I assume that intuitions about manipulation cases are critical to supporting premise 3 of the Argument from Manipulation, and return to the question of how else premise 3 might be defended in part 3.

There are two ways for compatibilists to respond to the Argument from Manipulation. The first compatibilist response, sometimes called the 'soft-line' response—the terminology is due to Kane (1985)—grants premise 3, but argues the conclusion is false because premise 1 fails. Rather, it is claimed that victims of manipulation fail to meet key conditions of freedom laid down by compatibilist views. The alternative 'hard-line' response grants premises 1 and 2, but denies premise 3. Instead, it is claimed that a manipulated agent who meets all CFP-conditions is indeed free. I take up each of these compatibilist responses to the Argument from Manipulation in turn.

1.2. The Compatibilist Soft-line Response

According to the compatibilist soft-line response, manipulation prevents an agent from realizing key CFP conditions. Let us examine two specific pathways by which manipulation might exert these effects—*corrupted information* and *deep self discordance*. Start with corrupted information. Consider the question of how a manipulator might achieve her goal of getting the 'target', i.e., the person being manipulated, to perform only certain very specific actions at specific times (and not other actions at other times). The philosophers who have advanced manipulation cases actually provide precious little information to help answer this question, and instead make only broad references to 'rigorous practices of conditioning' (Pereboom, 2001), vaguely Skinnerian kinds of behavioral engineering (Kane, 1996), and 'scripting' major life events (Greene and Cohen, 2004). But if we attempt to fill in the details about how manipulation might succeed, the most plausible scenarios involve *significant corrupting of the information used by the target as a basis to make decisions*. Consider the case of Plum above. How specifically

[4] See Pereboom (2001, pg. 112–115), Kane (1985, pg. 40), and Stump (2002, pg. 47–48).

might a Manipulator guarantee that young Plum will grow up to be an adult who will eventually kill Mrs. White? The answer would presumably have to involve some combination of *exclusively exposing* Plum to negative information about Mrs. White, *suppressing* Plum's hearing of opinions that dissent from this anti-Mrs. White propaganda, *selectively instilling* the idea that murder is the best way of achieving one's ends, *preventing* Plum from experiencing situations in which non-violent modes of conflict resolution are successful, *concealing* from Plum the horrific negative effects of killing, and so on…. But if manipulation works by corrupting the information available to the target in this way, then it is plausible that targets of manipulation will in turn be compromised in deploying various freedom-relevant psychological capacities. For example, in the case of Plum, adult Plum will have sharply limited and distorted information about the alternative ways of conducting his life and will be impaired in his ability to assign appropriate evaluative weights to these alternatives. Thus Plum will be impaired with respect to basic volitional and evaluative capacities that compatibilists (and incompatibilists as well) insist are critical to an agent's being free.[5]

A second, related, pathway by which manipulation might prevent an agent from realizing key CFP conditions is deep self discordance (see Haji, 1998, Shoeman, 1978 for related discussions). Manipulation cases operate on the assumption that environmental engineering can instill in the agent a *comprehensive* suite of the manipulator's chosen attitudes. But this possibility only makes sense if the victim's mind is something like a *tabula rasa*, and does not already have previously entrenched values, attitudes, and behavioral tendencies. But this tabula rasa assumption is deeply implausible, even when manipulation begins in the earliest days of infancy. More importantly for the present purposes, this *tabula rasa* assumption is surely unlikely to be *thought* to be true by most people.[6] It follows then that when presented with a manipulation case, many people will understand the target's post-manipulation

[5] The notion of *corrupted information* being developed here is importantly distinct from the notion of *insanity* (i.e., compromised normative competence) that figures heavily in the account of free will developed by Susan Wolf (Wolf, 1990, Wolf, 2003). For a helpful discussion of the differences between the two notions, see Faraci and Shoemaker (2010).

[6] The evidence is now overwhelming (see for example Harris (1998), Plomin et al (2001), and the section on 'Nativism' in Mason et al (2008)) that a sweeping anti-nativist view of the contents of the minds of human infants is untenable, and most educated people know this, either through direct contact with the evidence, or through participation in the cultural milieu in which this evidence is routinely conveyed. These individuals are likely to believe (correctly) that manipulation that begins even at the earliest days of infancy will of necessity overlay (or replace) significant elements of the agent's pre-existing evaluative repertoire.

psychology as involving two layers of attitudes. There is a surface layer that consists of the superficial values, attitudes, and behavioral tendencies instilled by manipulation. And there is a deeper layer in which the agent's antecedently present 'real' values, attitude, and core commitments reside. Moreover, the attitudes that reside in the agent's deep and surface layers may not be in harmony. For example, in the case of Plum, people may understand his adult psychology as involving substantial discordance between surface attitudes instilled by manipulation that endorse killing Mrs. White, and deep attitudes that lie below that abhor killing (and which Plum shares with most all normal humans). Of note, the notion of an agent's deep self, and the related ideas that an agent can fail to be identified with, or can be actively alienated from, her own attitudes has been developed in sophisticated ways in recent philosophical writings (Watson, 1975, Wolf, 1990, Frankfurt, 1988). My purpose here has been to sketch only the simplest and least theoretically committed version of a deep versus surface distinction. Though simple, this distinction is sufficient to capture in the broadest terms one pathway by which manipulation might undermine key CFP conditions.[7]

Notice that people's tendency to understand a manipulated agent as being in a state of deep self discordance and their tendency to see the agent as a victim of corrupted information need not be independent, and indeed are likely to be intimately connected. Recall that an agent who suffers from corrupted information will deliberate among a narrowed set of options and assign inappropriate weights to whatever options she envisions. Hence, she will end up selecting courses of action that do not reflect, and indeed might even actively discord with, her underlying values and commitments (i.e., the agent will be in a state of deep self discordance). Thus these two pathways by which manipulation might prevent an agent from realizing key CFP conditions—corrupted information and deep self discordance—are in fact likely to be closely connected, with people's judgments that the agent suffers from the former supporting judgments that she also suffers from the latter.

1.3. The Compatibilist Hard-line Response

The compatibilist soft-line response says that a manipulated agent does not in fact fulfill key CFP conditions and therefore is not free. The compatibilist hard-line response, in contrast, says the manipulated

[7] There are close cousins of manipulation arguments, such as Alfred Mele's intriguing zygote argument (Mele, 2006, pg. 184–195), in which it is unambiguous that the manipulator *creates* all the target's mental states, without rewriting or overlaying any existing mental states. I believe these 'creation' cases, in contrast to manipulation cases, do not generate an intuition that the agent is unfree, and I expand on this point elsewhere (Sripada, in preparation).

agent *does* fulfill all CFP conditions and therefore *is* free (i.e., this response grants premises 1 and 2 of the Argument from Manipulation, but claims that premise 3 is false). Compatibilists diagnose intuitions in manipulation cases as arising from our tendency to see a manipulated agent as exhibiting deficits in psychological capacities. But it is a corollary of this compatibilist position that if we do *not* envision the agent as suffering from psychological deficits (that is, if we envision the agent as having been spared any kind of psychological damage), then a hard-line response is justified—the manipulated agent is in fact free.[8] For example, Harry Frankfurt discusses two ways that manipulation might operate. One possibility involves ongoing intrusive modifications of the agent's psychology, which Frankfurt says clearly renders the agent unfree. Referring to the Manipulator as the 'D/n' for 'Devil/neurologist', Frankfurt then writes,

> The other possibility is that the D/n provides his subject with a stable character or program, which he does not thereafter alter too frequently or at all, and that the subsequent mental and physical responses of the subject to his external and internal environments are determined by this program rather than by further intervention on the part of the D/n. In that case there is no reason … against allowing that the subject may act freely or against regarding him as capable of being morally responsible for what he does.' (Locke and Frankfurt (1975), pg. 27; see also Kapitan (2000), Hajji (1998), Mckenna (2008) for related arguments along these lines)).

Of course, many compatibilists believe that envisioning the agent in a manipulation case as not suffering from psychological deficits is *difficult* because, as discussed above, the most plausible pathways by which manipulation might succeed appear to inevitably inflict psychological damage on the agent. But to the extent that we *are* able to imagine the manipulated agent as having been spared psychological damage, compatibilists often concede that we would then judge that the agent is free. Putting together these soft-line and hard-line strands in compatibilist thinking yields the following full statement of the compatibilist position:

---

[8] This hard-line response presupposes that the idea of a manipulated agent who is psychologically undamaged is in fact *coherent*, which many compatibilists either implicitly or explicitly grant. For example, Michael McKenna writes, '… I can see no way to foreclose the metaphysical possibility that the causes figuring in the creation of a determined morally responsible agent could not be artificially fabricated' (McKenna, 2008).

'Compatibilist Position' Regarding Intuitions in Manipulation Cases

1. Intuitions in manipulation cases are sensitive to whether or not there is damage to the manipulated agent's psychological capacities (due to factors such as corrupted information and deep self discordance).

2. (Soft-line) To the extent that the manipulated agent *is* seen as exhibiting damaged psychological capacities, the agent is intuitively judged to be unfree.

3. (Hard-line) To the extent that the manipulated agent is *not* seen as exhibiting damaged psychological capacities, the agent is intuitively judged to be free.

2. Intuitions about Manipulation: An Experimental Investigation

The Compatibilist Position stated above readily lends itself to empirical testing. To see this, it is important to recognize that at its heart, the Compatibilist Position is making a claim about the association between two judgment variables: 1) judgments that the manipulated agent suffers from psychological impairments; and 2) judgments that the manipulated agent lacks free will. The psychological sciences have developed a sophisticated toolkit of techniques for assessing the interrelationships between judgment variables including questionnaire-based methods for measuring people's intuitive judgments as well as validated statistical methods for quantifying the strength of the relationships between these variables and the confidence that is warranted that hypothesized relationships between variables really exist.

To test the Compatibilist Position regarding intuitions in manipulation cases, I conducted two studies involving a total of 360 subjects.

Study#1

In Study#1, all subjects (n = 240) read the following *base vignette*.

Bill and Dr. Z[9]

One day, Bill sees a woman named Mrs. White as she is jogging in the park. Bill hates this woman, and deliberates about

[9] This vignette is loosely based on the Professor Plum vignette (presented earlier) from Pereboom (2001) and the 'Mr. Puppet' vignette from Greene and Cohen (2004).

what to do. After weighing his options, Bill decides he should kill her. Bill's mind is not clouded by rage or other extreme emotions. Rather, Bill thinks clearly and carefully about his own desires and values, and only then makes a decision. After he kills Mrs. White, Bill reflects on his action. He wholeheartedly endorses what he has done.

But there is more you need to know about Bill, and how he came to be the person that he is now:

There is a man named Dr. Z who is a scientific genius and who is an expert at indoctrination. Dr. Z hates Mrs. White and formed the following plan. Dr. Z would take an infant from an orphanage and raise the child himself. He would teach and reward just the right behaviors in the child so the child would hate Mrs. White and want her dead. He would script all the major events in the child's life to nurture and cultivate in the child the goal of doing whatever it would take to kill Mrs. White. Dr. Z tried this plan previously on five other children, and each time the child grew up to kill Dr. Z's intended targets.

Half of the subjects were in the *Manipulation* condition and read the following conclusion to the vignette:

Dr. Z implemented his plan for Bill. He took Bill from an orphanage when Bill was an infant. The plan worked—once Bill had grown up, Bill had the desire to do whatever it takes to kill Mrs. White. Dr. Z's plan was kept completely hidden from Bill. Bill never knew that Dr. Z implemented the plan.

The other half of the subjects were in the *No Manipulation* condition and read the following alternative conclusion to the vignette:

Dr. Z was getting ready to implement his plan for Bill. He was about to take Bill from an orphanage when Bill was an infant. But at the last minute Bill was adopted by another family. But completely by chance, it turned out that Bill came to hate Mrs. White without any influence from Dr. Z at all. Once Bill had grown up, Bill had the desire to do whatever it takes to kill Mrs. White. Thus Bill turned out exactly how Dr. Z planned all along, but Dr. Z did not actually implement his plan at all.

Table 1:
Questions Used in this Study.

Question/ Variable #	Question Wording	Anchors for 7-point scale	Factor Name
1	How much do you agree with the statement: Bill killed Mrs. White of his own free will.	Strongly Agree, Strongly Disagree	Free Will
2	How much do you agree with the statement: Bill was in control of whether or not he killed Mrs. White.	Strongly Agree, Strongly Disagree	
3	How much do you agree with the statement: Bill is morally responsible for killing Mrs. White.	Strongly Agree, Strongly Disagree	
4	How much do you agree with the statement: Bill killed Mrs. White based on false information about her, and he was deprived of any opportunity to learn the truth.	Strongly Agree, Strongly Disagree	Corrupted Information
5	How much do you agree with the statement: Bill was never taught about why certain actions are right and wrong, so he does not truly know that killing Mrs. White is wrong.	Strongly Agree, Strongly Disagree	
6	How much do you agree with the statement: Bill killed Mrs. White because his upbringing kept him ignorant of alternative, non-violent, ways of acting.	Strongly Agree, Strongly Disagree	
7	How much do you agree with the statement: Bill's killing of Mrs. White does not reflect the kind of person who he truly is deep down inside.	Strongly Agree, Strongly Disagree	Deep Self Discordance
8	How much do you agree with the statement: The real Bill did not truly want to kill Mrs. White—Bill killed only because Dr. Z wanted him to.	Strongly Agree, Strongly Disagree	
9	How much do you agree with the statement: Bill is constrained by Dr. Z to act in a way that differs from how he himself, deep down, wants to act.	Strongly Agree, Strongly Disagree	

Subjects then answered a number of questions (Table 1) designed to probe their perceptions of Bill's level of free will, and the degree to which he suffers from corrupted information and deep self discordance.

Results showed that the two different versions of the vignette produced the expected directional effects on people's intuitive judgments.[10] Subjects presented with the Manipulation version of the vignette were significantly less willing to say that Bill killed Mrs. White of his own free will (Q1), was in control of whether or not he

[10] See Table A1 in the appendix for complete reporting of results for all questions in this study.

killed Mrs. White (Q2), and was responsible for killing Mrs. White (Q3), compared to subjects presented with the No Manipulation version of the vignette (all p values < 0.001). For example, 91% of subjects in the No Manipulation condition expressed agreement[11] with the statement 'Bill killed Mrs. White of his own free will' compared to only 36% of subjects in the Manipulation condition, and similar differences were observed for the other two free will-related questions (Q2 and Q3). These results appear to support the position that the folk as a group share philosophers' intuitions that manipulation compromises free will. But the decisive question (i.e. the question that divides incompatibilists and compatibilists) remains: What features of cases involving manipulation are people's free will intuitions responding to? To address this question, I conducted a series of further analyses.

Correlation Analysis

The Compatibilist Position predicts that there will be an inverse relationship between ratings of whether Bill killed Mrs. White of his own free will (Q1; 'free will ratings') and ratings of whether Bill suffered from corrupted information (Q 4-6) and deep self discordance (Q 7-9). *Correlation analysis*, which measures the association between two variables, provides a useful means to test this prediction. Results strongly supported the compatibilist view, with large and highly statistically significant negative correlations observed between free will ratings and ratings for corrupted information (Q 4-6) and deep self discordance (Q 7-9) (correlations ranged from -0.436 to -0.650; all p values < 0.001; see Table A2). Put another way, there was a very strong (and highly statistically significant) tendency for subjects who saw Bill as suffering from corrupted information and deep self discordance to also see Bill as lacking free will.

Simple Mediation Analysis

The Compatibilist Position makes a still stronger prediction. It predicts that when subjects are presented with two cases, matched in all respects but for the fact that one case involves manipulation but the other does not, then the *difference* between the two cases in people's judgments about free will should be mediated by the *difference* between the two cases in people's judgments about the presence of corrupted information and deep self discordance. *Mediation analysis*

[11] Agreement was assessed as a score of 1, 2, or 3 on a 7 point scale where 1 = Strongly agree and 7 = Strongly disagree.

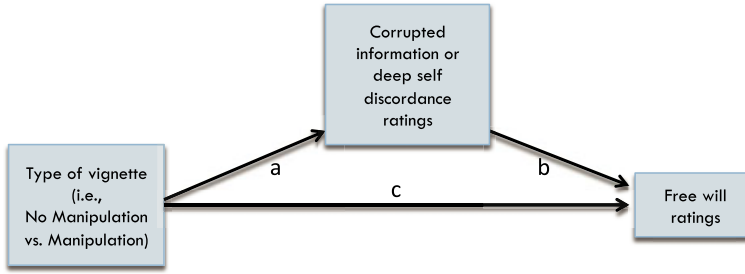

Figure 1: Simple Mediation Model. In this model, the type of vignette presented to subjects (No Manipulation versus Manipulation) affects free will ratings through either a direct path (path c), or an indirect path (paths a and b) that involves corrupted information or deep self discordance ratings as mediators.

is well suited to test this compatibilist prediction.[12] Mediation analysis assesses the hypothesis that one variable (the predictor) affects another variable (the outcome) by influencing an intervening variable (the mediator) (Baron and Kenny, 1986). Figure 1 shows a mediation model in which the vignette type variable (i.e., presenting subjects with either the No Manipulation or Manipulation version of the vignette) potentially influences free will judgments by two pathways: 1) An indirect pathway (paths *a* and *b*) that involves corrupted information or deep self discordance ratings as mediators; or 2) a direct pathway (path *c*) that does not involve corrupted information or deep self discordance ratings as mediators. Compatibilists predict the indirect pathway will be statistically significant (since they claim that free will judgments in manipulation cases are driven by judgments that the agent suffers from corrupted information and/or deep self discordance). Six separate mediation models were constructed, one for each of the six candidate mediating variables (i.e., Q 4-9). In five of the six models (all models except for Q6), the indirect pathway was found to be highly statistically significant (all p values < 0.001). Moreover, the mediating variables individually accounted for a sizable portion (range between 10–45%) of the variance in free will judgments.[13] These results provide strong evidence that the difference in free will judgments between the two versions of the vignette was significantly influenced by people's judgments

[12] In an insighful study, Nahmias and Murray (2010) used mediation analysis to probe how the folk understand (or misunderstand) agency in a deterministic world with resulting implications for free will.

[13] See Table A2 for the results of the six mediation models.

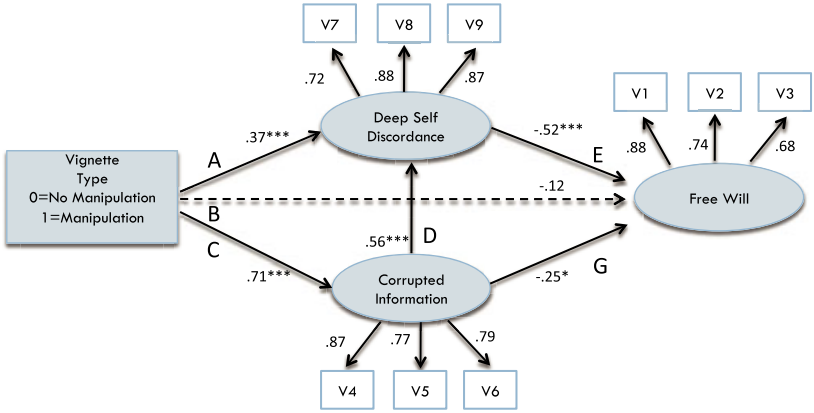

Figure 2: Structural Equation Model. In this model, two candidate mediators (Deep Self Discordance and Corrupted Information) fully explain the relationship between the type of vignette presented and Free Will judgments. Statistically significant paths are shown as solid arrows, while non-significant paths are shown as dashed arrows. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

about corrupted information and deep self discordance, and thus they support the Compatibilist Position.

Structural Equation Modeling

Structural equation modeling (SEM) provides a still more powerful way to test the Compatibilist Position.[14] Figure 2 shows a structural equation model of the results of this study. The model is the end product of a sequence of steps in which the assumptions of the model were tested and the fit of the overall model was evaluated. These steps are somewhat technical and described in the Appendix. In short, the model passed the preceding 'quality checking' steps. Here, my aim is to convey an intuitive sense of the findings from the model. Readers interested in the details are referred to the Appendix.

[14] SEM is an established statistical approach in the behavioral sciences (PsychInfo retrieves > 10,000 citations for SEM over the last 3 decades; see Kline, 2005 for an introduction to SEM). SEM allows several capabilities that go beyond simple mediation analysis. These include: 1) The ability to model *measurement relationships* in which a group of observed variables serve to jointly indicate an underlying unobserved variable; 2) The ability to model complex interrelationships between variables (rather than just a simple triangular mediation relationship); and 3) the availability of *fit statistics* that allow for significance testing of the overall model, rather than just significance testing of individual paths within a model.

Let us begin with an orientation to the overall model. On the left hand side of the model is a rectangle that represents the Vignette Type variable, i.e., presenting to subjects either the No Manipulation or Manipulation version of the vignette. This variable potentially influences two candidate mediators, the Corrupted Information variable and Deep Self Discordance variable (shown in the middle of the model), which in turn potentially influence the outcome variable, the Free Will variable (shown on the right hand side of the model). Notice the candidate mediating variables and the outcome variable are shown as circles, and are connected by arrows with a number of other variables shown as unshaded rectangles. The circles are *latent variables* (also called 'factors') and unshaded rectangles are *indicator variables*. The values of the latent variables are not directly observed, but rather their values are estimated from the indicator variables, which *are* directly observed. The values of the indicator variables are the ratings obtained from the questions presented to subjects (Table 1), and they are grouped so that closely related questions serve to indicate a common underlying latent variable. In particular, questions 1–3 indicate an underlying Free Will factor, questions 4–6 indicate an underlying Corrupted Information factor, and questions 7–9 indicate an underlying Deep Self Discordance factor. The arrows linking one variable to another are *paths*, and the coefficients associated with paths represent the strength of the relationship between the two variables linked by the path. Path coefficients range from 0 to 1, with a coefficient of 1 indicating a perfectly predictive relationship between the two variables linked by the path, while a coefficient of 0 indicates the absence of a relationship between these two variables (specifically along that particular path).

In the SEM model shown in Figure 2, there are two critical findings. First, the coefficient of the *direct* path linking the Vignette Type variable to the Free Will variable is *not* statistically significant. The second important finding concerns the three *indirect* pathways linking the Vignette Type variable to the Free Will variable (paths *A&E*, paths *C&G*, and paths *C,D&E*). Notice the coefficients along these three indirect paths are *all* statistically significant. The combination of these two findings provides very strong evidence for the Compatibilist Position.

To see why, recall that compatibilists believe that people's intuitions about free will in manipulation cases are driven by judgments of the presence of deep self discordance and corrupted information. Thus, compatibilists would predict that the paths along the three indirect paths that involve Corrupted Information and Deep Self Discordance as mediators will be large, which is in fact exactly what was

found. Indeed, formal mediation calculations reveal that 82% of the effect of the Vignette Type variable on Free Will judgments is accounted for by these indirect paths. The fact that the direct path is *not* statistically significant is also important. This means that the indirect paths involving Corrupted Information and Deep Self Discordance as mediators *fully* account for the influence of the Vignette Type variable on Free Will judgments. Put another way, reading the No Manipulation versus Manipulation version of the vignette had no significant effect at all on Free Will judgments over and above the effect it had on producing different interpretations of whether the scenario involved Corrupted Information and Deep Self Discordance. Thus, once the influence of Corrupted Information and Deep Self Discordance is accounted for, there is no *further* difference in Free Will judgments between the Manipulation and No Manipulation conditions of this study.[15]

Study#2: The Effect of Providing More Information about Manipulation

A sizable number of subjects in Study#1 saw manipulation as inflicting psychological damage on the agent.[16] One worry about this result is that perhaps subjects perceived the agent as suffering from psychological damage because they *misunderstand* how manipulation is supposed to work. Manipulation is supposed to be non-coercive in that it operates *through* an agent's desires and values rather than in opposition to them. Moreover, the upbringing of a manipulated agent is supposed to resemble a normal upbringing in most all respects, and is not supposed to involve complete ignorance of morality or other kinds of enforced ignorance. Now it may very well be *difficult* for subjects to imagine manipulation cases in this way. After all, as noted in Section 1, it is quite difficult to see how manipulation might succeed without inflicting damage to the agent's psychology. Nonetheless, to minimize

[15] Incompatibilists, on the other hand, would predict that *only* the direct pathway will be statistically significant and the indirect pathways will not. This is because they believe that intuitions about free will are not influenced by deep self discordance and corrupted information, but rather are influenced by another factor (i.e., absence of ultimate control) that is not explicitly measured in the current study. Any unmeasured mediators of the relationship between the Vignette Type variable and Free Will judgments would contribute to the overall effect along the direct path, which is why incompatibilists should predict that the direct path will be large. However, the coefficient along the direct path is in fact not statistically different than 0, so this prediction clearly was *not* supported by the data.

[16] Among subjects in the Manipulation condition in Study#1, questions about Deep Self Discordance yielded an average of 38% agreement and questions about Corrupted Information yielded an average of 63% agreement, with the remainder neutral or disagreeing.

the possibility that subjects are misunderstanding how manipulation is supposed to work, it would be helpful to utilize a vignette that provides even clearer descriptions of the manipulated agent's upbringing and how manipulation impacts the agent's adult psychology. For this reason, I conducted a second study. Subjects (n = 120) in this study read the same vignette from the Manipulation condition of Study#1. At the end of this vignette, they read the following additional paragraph.

> Bill is like anyone else in many respects. As he was growing up, Bill was educated about morality, the difference between right and wrong, and various ways he might conduct his life. Additionally, Bill was not simply fed lies about Mrs. White—he knows the truth about who she is and he knows exactly why he dislikes her. Bill is not a robot who simply does as others instruct. Nor is he under the grip of an irresistible impulse. Rather, Bill is a person, with desires, values, hopes, and dreams just like anyone else. But Bill's desires include killing Mrs. White. And his core values permit killing Mrs. White. So that is exactly what he does.

Subjects then answered the same questions as in Study#1. It was hypothesized that this additional paragraph of information about how manipulation is supposed to work would lead subjects to be *less* likely to judge the agent suffers from corrupted information and deep self discordance. However, if this is correct, then the Compatibilist Position implies two further predictions. First, it predicts that to the extent that Bill is judged *not* to suffer psychological damage, then Bill will be *more* likely to be judged to be free. Second, the Compatibilist Position makes the same mediation prediction as in Study#1. That is, it predicts that differences in free will judgments between the 'High Information' Manipulation vignette from Study#2 and the No Manipulation vignette from Study#1 will be fully explained by corrupted information and deep self discordance judgments as mediating variables.

Results from Study#2 showed that the addition of the extra paragraph of information had its intended directional effect. Compared to subjects in the Manipulation condition from Study#1, subjects in Study#2 expressed significantly *less* agreement with statements that Bill suffered from corrupted information (i.e., Q4, Q5, Q6; all p values < 0.001). They also expressed less agreement with statements that Bill suffered from deep self discordance (i.e., Q7, Q8, Q9), though only for Q8 did the differences reach statistical significance ($p < 0.05$).

Prediction#1 was confirmed in that in response to the Study#2 Manipulation vignette, the majority of subjects (66%) judged that Bill *does* have free will in killing Mrs. White. Statistical testing revealed that this percentage is significantly *higher* in Study#2 compared to the Manipulation condition from Study#1[17] (though significantly *lower* than the No Manipulation condition from Study#1[18]).

Simple mediation analysis was used to test Prediction#2. Results showed that judgments that Bill suffered from corrupted information (specifically Q4) and deep self discordance (specifically Q7, Q8, and Q9) played a highly significant role in explaining people's free will judgments (p values for the mediation pathways for these variables were all < 0.001; see Table A4). Moreover, once differences in deep self discordance judgments were controlled for, there was no longer *any* difference in free will judgments between the Manipulation vignette in Study#2 and the No Manipulation vignette in Study#1 (Table A4).[19] In other words, similar to the results from Study#1, reading the Manipulation vignette from Study#2 had no significant effect at all on free will ratings over and above the effect it had on producing judgments that the agent suffered from certain kinds of psychological damage—just as the Compatibilist Position predicts.

3. Implications for the Philosophical Debate about Manipulation

The preceding studies used correlation analysis, simple mediation analysis, and structural equation modeling to show that when presented with manipulation cases, subjects' judgments that the agent is unfree were strongly predicted by their judgments that the agent suffered from corrupted information and deep self discordance. Indeed, these two factors *fully* explained variation in people's free will judgments so that once variation in these two factors was accounted for, there were no remaining differences in free will judgments between vignettes involving manipulation and a matched vignette in which manipulation was absent. This finding supports the conclusion that our

---

17 66% vs. 36%; $X^2 = 21.61$, $p < 0.001$.

18 66% vs. 91%; $X^2 = 22.10$, $p < 0.001$.

19 A single mediation model was constructed in which Q7, Q8, and Q9 served as mediators. This model accounted for 90% of the variation in free will judgments, and the direct pathway was no longer statistically significant ($p = 0.499$). A full structural equation model of differences in responses to the Manipulation vignette in Study#2 and the No Manipulation vignette in Study#1 was not attempted. This is because several variables differed between the two vignettes either only minimally (Q2 and Q3 differed by less than 1 point on a seven point scale), or not at all (Q5).

intuitions in manipulation cases track the features that compatibilists say they track (i.e., the agent's possessing impaired psychological capacities) and do not track the features that incompatibilists say they track (i.e., the agent's lack of ultimate control over his or her actions).

The results of this study have implications for the debate between incompatibilists and compatibilists over free will. Incompatibilists claim that compatibilists cannot accommodate our intuitions that a manipulated agent is unfree. The results of this study suggest that this charge against compatibilism is incorrect. These results instead suggest that the intuition that a manipulated agent is unfree is driven primarily by judgments that the manipulated agent is compromised with respect to precisely the kinds of psychological capacities compatibilists regard as the basis for free will. Indeed, the results of the present study suggest, somewhat ironically, that far from being a *problem* for compatibilism, manipulation cases might even be seen as a kind of *justification* for compatibilist views. Insofar as what drives folk intuitions in manipulation cases is subtle tracking of just the kind of freedom-conferring conditions that compatibilists have long defended, then it seems that manipulation cases provide *evidence* for compatibilist principles by showing these principles are indeed deeply enshrined in the folk conception of freedom.

Incompatibilists can respond to some of these claimed implications of this study in at least two ways. One incompatibilist response challenges that this study accurately characterizes intuitions in manipulation cases (or more specifically, that this study accurately characterizes what features of manipulation cases our intuitions are responding to). The second line of response concedes that this study accurately characterizes intuitions about manipulation cases (or at least does not seek to press an objection along these lines), and instead denies that intuitions about manipulation cases matter to the success of the Argument from Manipulation. This line of response thus sees the present study as for the most part irrelevant to the philosophical debate about manipulation and free will. I take each of these responses up in turn.

The first incompatibilist response claims that this study fails to accurately characterize what features of manipulation cases our intuitions are responding to. One approach to pressing this objection focuses on the fact that the study was conducted with ordinary people who lack philosophical training or expertise. Manipulation cases typically require complex and sophisticated imaginings on the part of the subject whose intuitions about the case are being elicited. In particular, they require subjects to imagine and appreciate certain highly abstract and nuanced ideas, such as the idea that manipulation is *non-constraining* and *com-*

prehensive, and is supposed to operate *through* the agent's motivation rather than in opposition to it. If the philosophically naive subjects in this study thought that a manipulated agent suffers from corrupted information and deep self discordance, then, so says the objector, clearly they were not imagining the case correctly. In which case, why should their intuitions matter much to the philosophical debate?

In answering this 'the folk are too unsophisticated' objection, I begin by noting that the objection is only partially right—many subjects presented with the Manipulation versions of the vignette *did* perceive Bill as suffering from deep self discordance and corrupted information. But it is also important to realize that there was substantial variability in these judgments, and a sizable number of subjects *did not* perceive Bill as suffering from these problems. Indeed in Study#2, which was particularly explicit and detailed in characterizing Bill's development and adult psychology, *only a minority* of subjects regarded Bill as suffering from these problems.[20] Hence, in this respect, most subjects were imagining the case just as incompatibilists say one must. But what is most striking is not merely the presence of this variability, but rather how this variability was related to judgments about whether the agent is free: To the extent that people thought that Bill suffered from deep self discordance or corrupted information, or both, they tended to think Bill *lacked* free will in killing Mrs. White. While to the extent that people thought Bill did *not* suffer these afflictions, they tended to think that Bill *possessed* free will in killing Mrs. White. Hence, the objection fails because ordinary people *can* imagine manipulation cases as *not* involving impairments in psychological capacities, just as incompatibilists insist the case must be imagined. But the problem for incompatibilists is that to the extent that people imagined these manipulation cases as *not* involving psychological impairments in just the way that incompatibilists say they should, they commensurately *did not* see the agent as lacking free will. Rather, they tended to see the agent as free, which is precisely the pattern predicted by the Compatibilist Position as outlined in section 1.3.

A still deeper problem with the 'the folk are too unsophisticated' objection is that subjects in this study do *not* display a strange or incomprehensible pattern of responses. Rather, subjects evidence precisely the pattern of intuitive responses predicted by the Compatibilist Position. That is, subjects seem to respond to precisely those features that compatibilists have proposed are driving intuitions in manipulation cases (i.e., the presence of psychological impairments), and they respond in the

[20] Among subjects in the Manipulation condition in Study#1, questions about Deep Self Discordance yielded an average of 38% agreement and questions about Corrupted Information yielded an average of 63% agreement, with the remainder neutral or disagreeing. In Study#2, the respective percentages were 32% and 30%.

predicted way with judgments of diminished free will when they perceive such impairments as present. Moreover, the Compatibilist Position itself summarizes the views of leading compatibilist thinkers (see the Introduction and section 1). And there is no reason to believe these compatibilist thinkers are unsophisticated or that they fail to comprehend the subtleties of manipulation cases, and indeed quite the contrary. So given that the subjects in this study are exhibiting a pattern of responses that mimics the pattern defended by leading philosophical thinkers who *obviously* possess sufficient sophistication, then unless we are given some *specific* arguments otherwise, we should not be swayed by sweeping, generic charges that subjects in this study are too unsophisticated.

It is worth noting that I have undertaken several other studies to test folk intuitions about *ultimate control*, i.e., intuitions about whether free will and moral responsibility require a person to not only be able to shape her actions in light of her desires, values, and principles of choice, but also to be able to shape her most basic desires, values, and principles of choice themselves. Results of these studies show that people are quite willing to attribute free will and moral responsibility to an agent whose fundamental values and evaluative attitudes are *clearly and obviously* determined by factors that are completely out of her control (Sripada, in preparation). Collectively, these studies suggest a coherent pattern: people's intuitions appear *not* to be sensitive to the *origins* of an agent's basic evaluative attitudes (that is, it does not matter for free will if the agent has control over the initial formation of these basic evaluative attitudes; they can be instilled by biology, the environment, luck, or even a manipulator). But given that an agent already has a certain set of basic evaluative attitudes, people are very sensitive to the presence of obstacles or impairments that prevent an agent from selecting actions that reflect her own basic evaluative stance. I report the results of these studies of ultimate control in a separate paper, but I mention them briefly here to further dispel the 'the folk are too unsophisticated objection'. The fact that the folk are consistent and coherent in this way across studies provides additional evidence that the results of the present study are not simply due to the folk's lack of sophistication, but rather these results capture something genuine about the folk concept of free will.

A second kind of response by incompatibilists to the conclusions of this study concedes that this study accurately captures what features our intuitions respond to in manipulation cases (or at least it does not seek to press an objection on this issue), and instead seeks to deny that intuitions about manipulation cases matter much for the Argument from Manipulation. That is, the Argument from Manipulation, it is claimed, should succeed or fail based on principled argument, rather

than on whatever intuitions we might have about manipulation cases. While such a strategy is in principle available, it certainly is unlikely to be followed by most leading advocates of manipulation arguments such as Robert Kane and Derk Pereboom. These authors routinely make appeals to our intuitions in manipulation cases as part of their defenses of their positions,[21] thus evidencing the centrality of these intuitions to their own strategy for mounting manipulation-based defenses of incompatibilism.

It is actually hard to see how a defense of the Argument from Manipulation that completely eschewed the use of intuitions would work. Recall that intuitions about manipulation cases are supporting premise 3 of the argument, i.e., the claim that a manipulated agent is unfree. How might incompatibilists support this premise without recourse to intuitions about manipulation cases? The most obvious answer is that they would rely on any number of more general incompatibilist arguments, such as the Consequence Argument (Inwagen, 1983) or the Basic Argument (Strawson, 1994). The problem with this approach is that these arguments, because they are perfectly general, apply to manipulated agents and unmanipulated agents *equally* well. To the extent these general arguments are used to secure premise 3 of the Argument from Manipulation, the very generality of these arguments vitiates the topic of manipulation as being of any *distinctive* philosophical interest. Manipulation arguments are supposed to provide evidential weight in favor of incompatibilism that is *independent* of other arguments such as the Consequence Argument and Basic Argument. The most straightforward way for manipulation arguments to achieve this independence is to divorce them from these other more general incompatibilist arguments. One way to do this would be to grant premise 3 non-inferentially, that is, directly from intuitive judgments regarding manipulation cases. Alternatively, even if we do not outright grant premise 3 based on our intuitive reactions to manip-

[21] For example, Pereboom writes regarding his 'four cases' manipulation argument, 'The best explanation for the intuition that Plum is not morally responsible … is that his action results from a deterministic causal process…' (Pereboom, 2001, pg. 116). Kane writes '… at this level, the matter is one of ''introspection,'' or intuition. If a hard-core Hobbesian, or predestinationist, insists that CNC [covert nonconstraining] control does not take away from freedom in any significant sense, there is little one can say to argue him out of his position… Intuitively, most of us would concede that human freedom is freedom from every kind of control by others, coercive and non-coercive, overt and covert. Most modern compatibilists are likely to agree as well…'' (Kane, 1985, pg. 39). In a review of recent work on free will, Neil Levy and Michael McKenna (both compatibilists) write 'The Manipulation Argument relies upon our intuitive reactions to what is supposed to be objectionable [about] manipulation of an otherwise normally functioning agent' (Levy and McKenna, 2009, pg. 107).

ulation cases, it is hard to escape giving intuitions in manipulation cases at least significant evidential weight in whatever argument is eventually constructed.[22] These considerations lead me to believe that incompatibilists who want to mount a manipulation-based defense of their views have *no other option* than to accord intuitions about manipulation cases a central place in their arguments. It follows then that the results of the present study should be seen as quite problematic for incompatibilists, as these results suggest that these intuitions are actually responding to impairments in compatibilist psychological conditions for free will.

Before concluding, it is worth re-emphasizing several methodological points. This study illustrates how experimental methods, especially statistical techniques such as correlation analysis, mediation analysis, and structural equation modeling, can help philosophers solve the critical features problem—the problem of figuring out which among the myriad features present in hypothetical cases are the critical one(s) our intuitions are responding to. As I noted earlier, the debate between compatibilists and incompatibilists about manipulation cases is not about the *content* of intuitions in response to these cases. Both camps, for the most past, agree that there is a compelling intuition that manipulated agents are not free. The debate has rather been about what features of manipulation cases drive our intuitive responses. Given well-entrenched views on either side that our intuitions are responding to quite different things in these cases, prospects appear dim that armchair methods alone can break this impasse. Indeed, given extensive psychological evidence that people are generally quite unreliable about figuring out the factors that are driving their intuitive judgments[23] (see Sripada and Konrath (2011) for a full discussion of this point), one should have a healthy dose of skepticism towards claims that arise

---

[22] See for example the exchange between Derk Pereboom and Michael McKenna regarding the Argument from Manipulation in which intuitions about manipulation cases play a key role for both thinkers (Mckenna, 2008, Pereboom, 2008, especially section 2). Both philosophers appear to agree that intuitions by themselves do not *decisively* answer whether a manipulated is free. Rather intuitions serve a number of less dispositive though nonetheless quite important purposes, such as justifying which attitudes it is initially rational to hold and altering burdens of proof.

[23] It is interesting to consider whether the subjects in this study whose intuitions were shown to be responding to judgments of corrupted information and deep self discordance were *aware* that these factors were what their intuitions were in fact tracking. A reasonable hunch is that they lacked this awareness. Rather, I suspect that if asked how they came to believe that Bill in the manipulation version of the vignette lacks free will, they might have cited superficially salient features of the case (such as that Bill's desires were caused to arise by another agent), even though there is good reason to believe that this feature is *not* what their intuitions in fact responded to.

exclusively from the armchair that purport to decisively answer the question of what features of hypothetical cases our intuitions are responding to. Thus, where philosophical debates hinge critically on questions about what features of cases our intuitions track, philosophers should avail themselves of additional empirical-cum-statistical methods that can complement existing *a priori* methods of philosophical inquiry.

Conclusion

Incompatibilists and most compatibilists agree that a manipulated agent, as described in a standard manipulation case, is unfree, though they differ about what features of these cases our intuitions are responding to. In this paper, I investigated this issue experimentally. Results showed that subjects' tendency to judge that a manipulated agent is unfree was fully explained by their judgments that the manipulated agent suffers from damage to certain key psychological capacities. These results strongly support the view that intuitions in manipulation cases are responsive to impairments in just the kinds of psychological capacities that compatibilists have long claimed are the basis for free will. They also put serious pressure on incompatibilists' use of manipulation cases to support their position.

Appendix

All subjects were recruited using Amazon's Mechanical Turk, a validated method of subject recruitment and data collection (Buhrmester et al., 2011). Subjects were directed to a secure website where they completed an online survey, and were then paid for their time. The questions used in this study were presented in a randomized order. Additional questions probed Dr. Z's responsibility, as well as Bill and Dr. Z's degree of blame and punishment, and were not part of the present analysis. All dependent variables were inspected for deviations from normality assumptions. For all variables, skewness and kurtosis were less than 1.5.

Simple mediation analysis was conducted according to the recommentations of Baron and Kenny (1986). The effect size of mediation paths are reported as percentages in the main text as well as in Table A2 and Table A4. These percentages represent *mediation effect ratios*, defined as: (the *indirect effect* of the predictor on the outcome)/(the *total effect* of the predictor on the outcome) (Shrout and Bolger, 2002).

Structural equation analyses were performed with *EQS* (Multivariate Software Inc, Encino CA), a commercially available software

package for implementing structural equation modeling. The covariance matrix analyzed is shown in Table A3. Due to multivariate non-normality (Mardia's normalized coefficient = 13.4), Satorra-Bentler adjustment was performed, and robust fit statistics are reported for all analyses. To verify indicators loaded on the hypothesized factors, a measurement model was tested (Figure A1). This model displayed good fit with the data ($S\text{-}BX^2(24, N = 240) = 33.3$, $p = 0.1$; NFI = 0.974; CFI = 0.993; RMSEA = 0.04, PCLOSE n.s.). Modification indices were thresholded at $p < 0.01$ to avoid capitalizing on chance relationships in the data (MacCallum et al., 1992), in accordance with practices recommended in the literature (Bentler, 1995). These indices did not recommend changing any relationships between indicators and latent factors. Alternative two factor measurement models (e.g., in which corrupted information and deep self discordance indicators are combined to indicate a single 'psychological impairment' factor) and a one factor model were tested and all models yielded substantially worse fit with the data than the three factor model shown in Figure A1. The structural equation model in Figure 2 in the main manuscript was selected based on antecedent theoretical hypotheses. This model displayed good fit with the data ($S\text{-}BX^2(30, N = 240) = 56.8$, $p = 0.002$; NFI = 0.965; CFI = 0.983; RMSEA = 0.06, PCLOSE n.s.), and modification indices did not recommend addition or removal of paths.

Table A1:
Results for Each Question for Study#1 and Study#2.

	Study #1		Study#2
Question/ Variable #	No Manipulation Condition Mean (sd)	Manipulation Condition Mean (sd)	'High Information' Manipulation Condition Mean (sd)
1	1.60 (1.2)	4.09 (1.8)	3.03 (1.7)
2	1.82 (1.6)	3.51 (1.8)	2.78 (1.7)
3	1.49 (1.2)	3.16 (1.8)	2.38 (1.5)
4	5.60 (1.6)	2.80 (1.8)	4.33 (2.1)
5	5.28 (1.7)	3.38 (1.8)	5.12 (1.8)
6	5.43 (1.6)	3.08 (1.7)	4.42 (1.9)
7	6.07 (1.3)	4.43 (1.7)	4.63 (1.7)
8	6.48 (1.1)	3.72 (1.8)	4.28 (1.9)
9	6.32 (1.1)	3.77 (1.6)	4.07 (1.8)

Table A2:
Results from Correlation Analysis and Simple Mediation Analysis for Study#1.

Question/ Variable #	Correlation with free will ratings (all p values < 0.001)	% of variance in free will ratings explained by mediation pathway	Statistical significance of mediation pathway
4	−0.577	30.0%	< 0.001
5	−0.527	22.2%	< 0.001
6	−0.436	10.1%	0.08
7	−0.476	17.2%	< 0.001
8	−0.650	44.7%	< 0.001
9	−0.593	33.1%	< 0.001

Table A3:
Covariance Matrix for Structural Equation Modeling Analysis for Study#1.

	Vignette	1	2	3	4	5	6	7	8	9
Vignette	0.251									
1	0.626	3.93								
2	0.425	2.546	3.63							
3	0.418	1.946	1.654	3.015						
4	−0.703	−2.526	−2.208	−1.906	4.872					
5	−0.475	−2.066	−1.671	−1.522	2.859	3.912				
6	−0.59	−1.769	−1.535	−1.513	3.054	2.616	4.188			
7	−0.412	−1.615	−1.226	−1.44	1.909	1.584	1.616	2.923		
8	−0.692	−2.609	−2.022	−1.789	2.797	2.29	2.185	2.211	4.095	
9	−0.64	−2.215	−1.701	−1.746	2.69	1.936	2.199	2.09	2.85	3.546

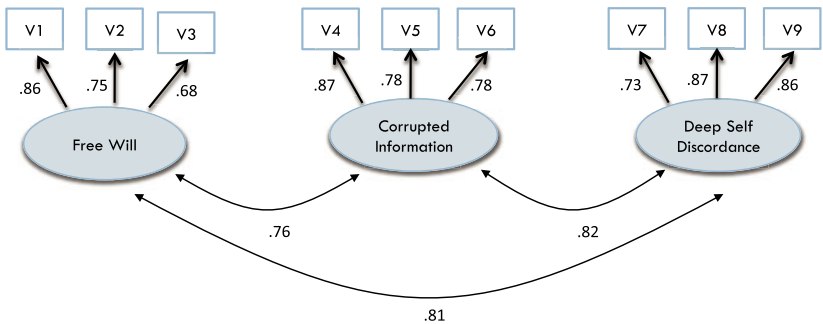

Figure A1: Measurement Model for Study#1. This model provides evidence that the ratings elicited from subjects (shown as boxes) serve to indicate three underlying factors (shown as ovals).

Table A4:
Results from Simple Mediation Analysis for Study#2. * = $p < 0.5$; n.s. = Not Significant.

Question/ Variable #	Correlation with free will ratings (all p values < 0.001, except where noted)	% of variance in free will ratings explained by mediation pathway	Statistical significance of mediation pathway
4	−0.329	15.8%	< 0.01
5	−0.157*	1.4%	n.s.
6	−0.237	8.2%	0.05
7	−0.584	48.07%	< 0.001
8	−0.606	70.3%	< 0.001
9	−0.585	71.3%	< 0.001
7,8,9		90.3%	< 0.001

References

Baron, R. and Kenny, D. 1986: The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*, 51, 1173–82.

Bentler, P. 1995: *EQS structural equation program manual*, Los Angeles: BMDP Statistical Software.

Buhrmester, M. D., Kwang, T. and Gosling, S. D. 2011: Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6, 3–5.

Dworkin, G. 1976: Autonomy and behavior control. *Hastings Center Report*, 6, 23–28.

Faraci, D. and Shoemaker, D. 2010: Insanity, Deep Selves, and Moral Responsibility: The Case of JoJo *Review of Philosophy and Psychology*, 1, 319–332.

Fischer, J. M. and Ravizza, M. 1998: *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

Frankfurt, H. 1971: Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, 68, 5–20.

——. 1988: Identification and externality. *The Importance of What We Care About*. New York: Cambridge University Press.

Gert, B. and Duggan, T. J. 1979: Free will as the ability to will. *Nous*, 13, 197–217.

Greene, J. and Cohen, J. 2004: For the law, neuroscience changes nothing and everything. *Philos Trans R Soc Lond B Biol Sci*, 359, 1775–85.

Greenspan, P. 1978: Behavior control and freedom of action. *The Philosophical Review*, 225–240.

Haji, I. 1998: *Moral Appraisability: Puzzles, Proposals, and Perplexities*, New York: Oxford University Press.
Harris, J. R. 1998: *The Nurture Assumption*, New York: Free Press.
Inwagen, P. V. 1983: *An essay on free will*, Oxford: Clarendon Press.
Kane, R. 1985: *Free Will and Values*, Alabany, NY: State University of New York Press.
—— 1996. *The Significance of Free Will*, Oxford: Oxford University Press.
Kapitan, T. 2000: Autonomy and Manipulated Freedom. *Philosophical Perspectives*, 14, 81–104.
Kline, R. B. 2005: *Principles and practice of structural equation modeling*, New York: Guilford Press.
Levy, N. and Mckenna, M. 2009: Recent Work on Free Will and Moral Responsibility. *Philosophy Compass*, 4, 96–133.
Locke, D. and Frankfurt, H. 1975: Three concepts of free action. *Proceedings of the Aristotelian Society*, *Supplementary Volumes*, 95–125.
Maccallum, R. C., Roznowski, M. and Necowitz, L. B. 1992: Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychol Bull.*, 111, 490–504.
Mason, K., Sripada, C. S. and Stich, S. 2008: The Philosophy of Psychology. In Moran, D. (ed) *The Routledge Companion to Twentieth-Century Philosophy*. London: Routledge Press.
Mckenna, M. 2008: A Hard-line Reply to Pereboom's Four-Case Manipulation Argument. *Philosophy and Phenomenological Research*, 77, 142–159.
Mele, A. 2006: *Free Will and Luck*, New York: Oxford Univeristy Press.
Nahmias, E. and Murray, D. 2010: Experimental philosophy on free will: An error theory for incompatibilist intuitions. In Aguilar, J., Buckareff, A. and Frankish K. (eds) *New Waves in Philosophy of Action*. New York: Palgrave-Macmillan.
Nisbett, R. and Wilson, T. 1977: Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84.
Pereboom, D. 2001: *Living Without Free Will*, New York: Cambridge University Press.
Pereboom, D. 2008: A Hard-line Reply to the Multiple-Case Manipulation Argument. *Philosophy and Phenomenological Research*, 77, 160–170.
Plomin, R., Defries, J. C., Mcclearn, G. E. and Mcguffin, P. 2001: *Behavioral Genetics*, New York: Worth Publishers.
Scanlon, T. M. 1986: The significance of choice. *The Tanner Lectures on Human Values*, 7, 149–216.

Shoemaker, D. 2003: Caring, Identification, and Agency. *Ethics*, 114, 88–118.
Shoeman, F. 1978: Responsibility and the problem of induced desires. *Philosophical Studies*, 34, 293–301.
Shrout, P. and Bolger, N. 2002: Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol Methods*, 7, 422–45.
Sripada, C. S. in preparation: Being the source of yourself: folk intuitions about free will and ultimate control.
Sripada, C. S. and Konrath, S. 2011: Telling more than we know about intentional action. *Mind & Language*, 26, 353–380.
Strawson, G. 1994: The Impossibility of Moral Responsibility. *Philosophical Studies*, 75, 5–24.
Stump, E. 2002: Control and Causal Determinism. In Buss, S. and Overton, L. (eds) *Contours of Agency: Essays on Themes from Harry Frankfurt*. Cambridge: MIT Press.
Vihvelin, K. 2007: Arguments for incompatibilism. In Zalta, E. N. (ed) *The Stanford Encyclopedia of Philosophy*.
Watson, G. 1975: Free Agency. *Journal of Philosophy*, 72, 205–220.
——. 1987: Free Action and Free Will. *Mind*, 96, 145–172.
Wolf, S. 1990: *Freedom within Reason*, Oxford: Oxford University Press.
——. 2003: Sanity and the metaphysics of responsibility. In Watson, G. (ed) *Free Will*. 2nd ed. Oxford: Oxford University Press.