Outsourcing the deep self:
Deep self discordance does not explain away intuitions in manipulation arguments

Gunnar Björnsson, Umeå University, University of Gothenburg

Manipulation arguments have figured prominently in recent debates about incompatibilism about moral responsibility, with compatibilists of various stripes seeing these arguments as particularly challenging.[1] In its basic form, a manipulation argument for incompatibilism tries to establish that there is a case of manipulation such that:

> UNDERMINING: It undermines the moral responsibility of the manipulated agent.
> EQUIVALENCE: It is not relevantly different from any other form of deterministic causal influence by factors outside an agent's control.

If both conditions are satisfied, any deterministic influence by factors outside an agent's control undermines moral responsibility.

In a recent paper, Chandra Sripada (2013) argues that compatibilists can explain the intuitions relied upon by manipulation arguments. Based on statistical analysis of responses to a representative manipulation case, Sripada concludes that manipulation doesn't undermine attributions of responsibility because it makes people see the manipulate agent's actions as determined by factors beyond the agent's control. Rather, it does so because manipulation undermines people's sense that the action in question reflects the agent's "deep self." And the requirement that responsible action concord with the agent's deep self, Sripada assumes, is a compatibilist one. If his argument is correct, the intuitions pumped by manipulation cases provide no support for incompatibilism.

Sripada's strategy is innovative and potentially helpful. In this paper, however, I identify weaknesses in Sripada's study and his interpretation of the data, and present new data strongly suggesting a very different picture. On this picture, most of the effect of manipulation on attributions of free will and moral responsibility is unmediated by worries about deep self discordance. Instead, such worries are largely accounted for by beliefs that the agent's action is ultimately explained by factors outside of the agent's control, and by resulting worries about the agent's free will.

If this picture is correct, it straightforwardly undermines Sripada's compatibilist account of manipulation worries, and supports the contention that manipulation cases bring forth implicit incompatibilist commitments. But notions of an agent's "deep self" have played important further roles in contemporary moral psychology beyond the use it is put to in

[1] Among the most widely discussed are arguments from Pereboom 2001 and Mele 2006; compatibilists who recognize the challenge provided by such arguments include Nelkin 2011, McKenna 2013 and Tognazzini 2014.

Sripada's argument. Many have argued that an agent is responsible for an action to the extent that it reflects what might be intuitively characterized as her deep, or true, or real self: her higher-order desires, or values, or plans, or "cares."[2] More recently, notions of a deep, true, or real self also play an increasingly prominent role among empirically oriented moral psychologists. Some have explored everyday judgments about an agent's deep self and proposed that such judgments affect judgments about a variety of other aspects of the agent, including the agent's happiness, values, weakness of will, responsibility, blameworthiness and praiseworthiness. For example, Sripada has argued that judgments of deep self concordance—about the match between an outcome and the agent's deep self—explain asymmetries in attributions of intentionality.[3] Relatedly, Thomas Nadelhoffer and colleagues (2013) argue that neuroscientific explanations of actions undermine folk attributions of responsibility for actions by making people think that the agent's deep self is insufficiently involved in bringing about the action. Following a different thread, George Newman and colleagues present evidence that people associate an agent's deep self more with her feelings than with her moral beliefs, and more with feelings or beliefs that they themselves approve of (Newman, Bloom and Knobe 2013). They also argue that when moral judgments affect people's attributions of happiness, values, weakness of will, or praise and blame, they do so by affecting beliefs about the attributee's deep self (Newman, Freitas and Knobe 2014). (Both papers by Newman and colleagues contain overviews of earlier empirical work in the area.)

If the picture suggested in this paper is correct, however, there is reason to suspect that judgments about what an agent "truly is," "truly wants," or "wants deep down" have a less fundamental explanatory role than some have thought. Moreover, the folk notion of the deep self has an importantly different extension than the philosophical notions. Where the latter concern some internal aspects of the agent's motivational structure, the former is partially influenced by thoughts about the source of that structure—the deep self is partially *outsourced*, we might say.

In what follows, I recount the core of Sripada's study and argument (section 1) and partially defend his general method against some natural objections, while raising specific worries about his study (section 2). I then present and analyze data from a study designed to mitigate those worries (section 3), and discuss the consequences of the resulting picture as well as ways that it might be resisted (sections 4 and 5).

[2] Among the most discussed modern deep self accounts of free agency and responsibility are Frankfurt 1971, Watson 1975, and Bratman 1997. For an influential early discussion, see Wolf 1990.

[3] See Sripada 2010 and Sripada and Konrath 2011. For criticism, see Rose et al. 2011, who argue that Sripada's use of structural equation modeling is flawed, and Cova and Naar 2012, who argue that Sripada's model makes the wrong predictions.

1. Sripada's study

The purpose of Sripda's study was to determine whether people are really committed to EQUIVALENCE, by determining why people take responsibility to be undermined in the relevant sort of manipulation case. Is it because they take certain compatibilist conditions for moral responsibility to be violated, or because they take the manipulated agent's behavior to be determined by factors outside his control, i.e. because of worries about conditions stressed by incompatibilist and skeptics about free will and moral responsibility? The traditional way that philosophers have approached such questions is through reflection on cases, but Sripada (2012: 566) suggests that this method is unreliable. One problem is that different people confronted with a given case might construct very different mental representations of that case, and so make judgments based on quite different grounds. Another is that people are generally bad at identifying the features of a situation on which their judgments are based.

Both problems can be mitigated using statistical methods. The questions at hand are in effect questions about relations of dependence between the values of different variables: between the presence or absence of manipulation, the degrees to which people think that someone is responsible, and the degrees to which they think that various putative compatibilist or incompatibilist conditions for responsibility obtain. In experimental sciences, one standard way of determining such dependence relations is to intervene on one variable and track how that affects the values of the other variables and their statistical or probabilistic relations. This is what Sripada did. He varied whether an agent in a scenario was manipulated or not, and looked at how this affected both attributions of moral responsibility and beliefs about whether putative compatibilist conditions for responsibility were satisfied.

Compatibilists have suggested a number of different such conditions, but Sripada focuses on two: conditions requiring that the agent is or could be aware the relevant facts of the situation, including moral facts, and conditions requiring that the action is in some sense expressive of the agent's "deep self," i.e. of who he truly is, or what he really wants. As Sripada (2012: 569–71) notes, standard manipulation cases in the literature can easily be understood as partly undermining these particular conditions, and Sripada hypothesized that this is what leads people to deny that the manipulated agent is responsible for his action, not the violation of some incompatibilist condition.

Here, then, are the core parts of Sripada's study: First, 240 subjects read the following story:

BILL AND DR. Z

> One day, Bill sees a woman named Mrs. White as she is jogging in the park. Bill hates this woman, and deliberates about 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 the subjects then read the MANIPULATED sequel to the story, and half the NOT MANIPULATED sequel:

MANIPULATION (MAN)

MANIPULATED: 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.

NOT MANIPULATED: 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.

Finally, all subjects were asked to indicate their level of agreement with the following three groups of statements, on a 7-point scale (labels: 'strongly agree', 'agree', 'somewhat agree', 'neither agree nor disagree', 'somewhat disagree', 'disagree', 'strongly disagree').

FREE WILL (FRW)

OWN FREE WILL (OFW): Bill killed Mrs. White of his own free will.
CONTROL (CTR): Bill was in control of whether or not he killed Mrs. White.
MORAL RESPONSIBILITY (MRR): Bill is morally responsible for killing Mrs. White.

CORRUPTED INFORMATION (CIN)

FALSE INFORMATION (FIN): Bill killed Mrs. White based on false information about her, and he was deprived of any opportunity to learn the truth.

MORAL IGNORANCE (MIG): Bill was never taught about why certain actions are right and wrong, so he does not truly know that killing Mrs. White is wrong.

PRACTICAL IGNORANCE (PIG): Bill killed Mrs. White because his upbringing kept him ignorant of alternative, non-violent, ways of acting.

DEEP SELF DISCORDANCE (DSD)

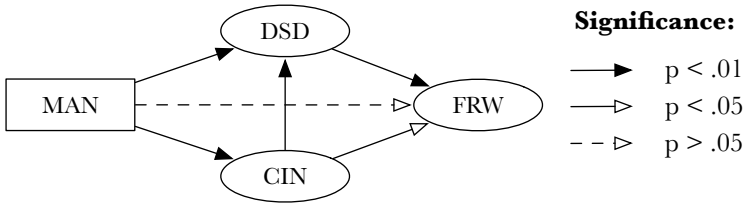
PERSON DEEP DOWN (PDD): Bill's killing of Mrs. White does not reflect the kind of person who he truly is deep down inside.

TRULY WANT (TRW): The real Bill did not truly want to kill Mrs. White—Bill killed only because Dr. Z wanted him to.

WANTED DEEP DOWN (WDD): Bill is constrained by Dr. Z to act in a way that differs from how he himself, deep down, wants to act.

Composite values were calculated for levels of agreement with FREE WILL (FRW), CORRUPTED INFORMATION (CIN) and DEEP SELF DISCORDANCE (DSD) statements. As expected, and as required for manipulation arguments to get off the ground, agreement with FREE WILL statements were significantly lower for subjects in the MANIPULATED condition. But Sripada's compatibilist hypothesis was that the effect of the MANIPULATION (MAN) variable on the FRW variable would be mediated by effects on variables designed to track violations of the measured compatibilist conditions on free will and moral responsibility, CIN and DSD. To decide this issue, Sripada explored a statistical model according to which the effect of MAN on FRW is either direct or mediated by scores on CIN and DSD, with possible interactions between CIN and DSD. In such a model, he found that MAN had significant indirect effects on FRW via DSD and CIN (accounting for 82% of the total effect of MAN on FRW), whereas the direct effect of MAN on FRW was weak and non-significant. (See Figure 1.) The model was found to fit the data well, seemingly strongly supporting Sripada's compatibilist-friendly explanation of why people take manipulation to undermine responsibility. (For statistical details, see Appendix A of Sripada 2012.)

Fig. 1 Sripada's model



2. Worries about Sripada's argument

A number of worries can be raised about Sripada's study and his claim that it undermines manipulation arguments. In this section, I will briefly discuss some of them; in the next, I present a study designed to address the more pressing of those pertaining specifically to Sripada's argument.

One general worry concerning Sripada's appeal to intuitions of laymen is that they might be less sensitive to some of the relevant distinctions than philosophers who have thought carefully about responsibility. This is obviously a valid worry, and we should be aware of the possibility that laymen fail to understand scenarios and related questions in the philosophically most relevant manner. However, it is not a worry that renders the consultation of laymen uninteresting. First, we can take steps to decrease risks of misinterpretation, as we will in section 3. Second, given that specialists tend to have conflicting intuitions and make conflicting judgments about crucial cases, and given the human propensity for confirmation bias, it is a live possibility that philosophers' intuitions are theoretically motivated. Perhaps compatibilists tend to focus on those aspects of the case that trigger positive intuitions of responsibility, whereas incompatibilists tend to focus on aspects that tend to trigger the opposite intuitions, each unconsciously playing down other aspects. (Fixing attention on certain aspects of a case is even part of McKenna's (2008: 144–5) explicit compatibilist strategy for responding to manipulation arguments.) To minimize this problem, systematic studies of intuitions of theoretically less committed individuals might provide a useful check. Third, when it comes to issues of moral responsibility, philosophers often purport to address concerns of considerable relevance outside the seminar room. For that reason, it might be particularly useful to understand the structure of everyday thinking about moral responsibility, and the extent to which incompatibilist worries can be seen as inherent in ordinary conceptions of responsibility.

Even if one agrees that studies of folk intuitions can be a useful complement to traditional philosophical methods, however, one might have specific worries about the relevance of subjects' responses in Sripada's studies. I now turn my attention to these.

The first worry is that intuitions about BILL AND DR. Z say little about how people understand the relation between moral responsibility and *determinism*, as nothing in the scenario ensures that Bill's behavior is completely determined by Dr. Z's manipulation. If this is right, the results from Sripada's study do not rule out that deterministic manipulation would have had a more of a direct effect on FREE WILL scores (cf. Gorin 2013). This is a legitimate worry, and something that one might want to address in further studies. But there are two reasons to think that the results would have been similar if the vignette had made it explicit that the manipulation was deterministic. The first reason is that Sripada's scenario both stresses Dr. Z's expertise in indoctrination and includes data on his previous repeated success with similar interventions. This, I think, provides reason to expect a sizable proportion of subjects to understand Bill's action as in fact determined by Dr. Z's manipulation. The other

reason is that even if the scenario does not explicitly state that Dr. Z *completely* determined Bill's actions, it seems to clearly convey that Dr. Z shaped Bill such that he would be very unlikely to refrain from killing Mrs. White. Because of this, Sripada's results might be taken to show that subjects do not take such strong probabilistic manipulation to undermine responsibility except through undermining compatibilist conditions. However, I would expect people who take deterministic causation by factors outside an agent's control to undermine responsibility to also take corresponding instances of strong probabilistic causation to do so, albeit to a lesser degree (cf. Mele's (2005) argument that indeterministic manipulation might undermine responsibility just as well as deterministic manipulation). For these reasons, I would not expect a study involving explicitly deterministic manipulation to support a radically different model of the interaction between MANIPULATION, FREE WILL, CORRUPTED INFORMATION and DEEP SELF DISCORDANCE.

The second specific worry I have in mind is that the FREE WILL statements used in Sripada's study fail to measure subjects' intuitions involving the relevant notions of control, free will and moral responsibility. For example, although *control* is often understood as a requirement of free will, incompatibilists are typically happy to recognize that determinism is compatible with *some* sort of control, in particular what Fischer (1994) calls "guidance control." (After all, we gladly say that thermostats "control the temperature.") A worry, then, is that even subjects with incompatibilist tendencies might attribute control to manipulated Bill, thus giving the FREE WILL measure a misleading compatibilist component. This worry would be easily handled by removing the CONTROL statement from the measure, but similar worries might arise with respect to free will and moral responsibility. Again, the worry would be that the kinds of free will and moral responsibility ascribed by subjects agreeing with OWN FREE WILL and MORAL RESPONSIBILITY statements are not of the variety that incompatibilists are concerned with, i.e. the kind related to what Pereboom (2001) calls "basic desert" of blame or punishment, i.e. desert not grounded in consequentialist or contractualist considerations, but instead the kinds that incompatibilists too often take to be compatible with determinism.[4] If this is right, what Sripada's study reveals is merely that when manipulation undermines these other forms of free will, control and moral responsibility, it does so primarily through undermining knowledge and deep self expressiveness. Attributions of basic desert might still, for all we know, be directly undermined by the sense that Bill's actions are ultimately determined by Dr. Z's intervention.[5]

[4] For discussion, see special issue of *Philosophical Explorations*, introduced by Pereboom and Sie (2013).

[5] For dissociations of different kinds of free will or responsibility, see e.g. Pereboom's (2015) thesis that although basic desert-entailing responsibility is incompatible with determinism, another important kind of moral responsibility is not, and John Fisher (e.g. 1994; 2002) claim that while one important sort of free will is incompatible with determinism, the sort of freedom or control required for responsibility is not.

Again, this seems to be a real possibility, and a possibility that further studies might want to minimize, for example by adding a statement about *desert* to the FREE WILL measure. However, it does not presently undermine Sripada's argument: a number of earlier studies of free will and moral responsibility suggest that attributions of desert of blame go hand in hand with attributions of moral responsibility (see e.g. Nahmias et al. 2007), and the presumption, until we have evidence to the contrary, should be that subjects who make judgments about moral responsibility and desert are concerned with what philosophers are debating.

The third specific worry is that other kinds of manipulation than that portrayed in Sripada's study are better suited to underpin manipulation arguments. Crucially, the manipulation involved needs to be one that both clearly undermines responsibility and does so because it is a factor determining (or nearly so) the agent's actions while being beyond the agent's control. More intrusive cases of manipulation tend to fare well with the first, UNDERMINING, requirement, but less well with the second, EQUIVALENCE, requirement. Less intrusive cases have the opposite problem. One might think that Sripada's case errs on the intrusive side, and that this explains why its responsibility undermining effect comes from violations of compatibilist conditions of responsibility.

This would be a serious problem if, as Sripada himself seems to think, the incompatibilists would predict that the statistically significant effect of MANIPULATION on FREE WILL would be direct *and only direct*, not mediated by DEEP SELF DISCORDANCE or CORRUPTED INFORMATION (2012: 580, n15). But part of the value of Sripada's statistical approach is that it provides a method of measuring whether responsibility scores are lowered *beyond* what is motivated by violation of compatibilist conditions. If they are, we should see a significant direct effect of MANIPULATION on FREE WILL. Since the mean agreement with the MORAL RESPONSIBILITY statement in the MANIPULATED condition was above the midline, the scale gave subjects with incompatibilist tendencies plenty of room to lower their responsibility scores beyond what was motivated by perceived violation of compatibilist conditions. The fact that Sripada did not see any significant direct effect can thus constitute evidence that subjects' FREE WILL answers were not affected by incompatibilist conditions. So intuitions about Sripada's case could still be appropriate for supporting a compatibilist account of manipulation-driven intuitions of undermined responsibility. There might be better cases on which to base a manipulation argument, but it would remain to show that Sripada's results would not generalize to those.[6]

[6] Relatedly, Gorin (2013) points out that because subjects do not take Bill to satisfy all standard compatibilist conditions, the BILL AND DR. Z case is relevantly unlike some of the cases appealed to by incompatibilists. Pereboom's (2001) Case 1, for example, is one where there *can* be no DEEP SELF DISCORDANCE because all the agent's values are instilled by the manipulators. However, Bill is also explicitly described as reflecting on his values and wholeheartedly endorsing his action, so any DEEP SELF DISCORDANCE attributed in this case would seems to be attributed in spite of the description of the case. If this is true about BILL AND DR. Z, it might be equally true about other cases, such as Pereboom's Case 1.

I do not claim to have conclusively addressed the worries mentioned thus far; in most cases that would require further studies. But I think that they are less pressing than problems pertaining to Sripada's DEEP SELF DISCORDANCE statements. The general problem here concerns the interpretation of the sort of statements involved. It is just not clear what it means to say that someone *truly* is a certain kind of person *deep down inside*, or acts in a way that differs from how he himself, *deep down*, wants to act, or performs an action that he did not *truly* want to perform. More specifically, there are two sorts of worries here, both of which might undermine Sripada's results: one concerning the multitude of conflicting springs of actions that might be identified with what an agent wants deep down, and one concerning the origins of such springs.

Consider first the worry about multiple conflicting springs of action. Since the values, preferences and desires of most people are highly complex, judgments about what someone truly wants, deep down inside, will have to be selective. It is not clear, however, what springs are selected when one makes such judgments. Perhaps one has in mind values that are best integrated with other behavioral dispositions, or most stably action-guiding under normal circumstances. But it seems likely that pragmatic or normative reasons also guide the selection. Those disposed to blame an agent might focus on aspects of his motivational setup that makes sense of blame, while those disposed to excuse him might think of his negative aspects as somehow external to him (cf. evidence that subjects' own values affect what they understand as the "true self" of an agent from (Newman, Bloom and Knobe 2014)). If agreement with DEEP SELF DISCORDANCE statements is affected by these sorts of pragmatic factors, one might worry that it results from, rather than explains, the sense that Bill's free will and responsibility are undermined by Dr. Z's manipulation.

The most straightforward way to test for this would be to directly manipulate DEEP SELF DISCORDANCE judgments and track whether this would change FREE WILL attributions. This is also what Sripada attempted. In a second study using only the MANIPULATED condition, he appended the following passage to the vignette, detailing how Bill had all relevant information and how his actions were in line with his desires and values:

> 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 (Sripada 2012: 581).

Compared to the MANIPULATED condition of the first study, this did indeed lower attributions of DEEP SELF DISCORDANCE while raising attributions of FREE WILL. Moreover, the increase in the latter was statistically mediated by the decrease in the former (ibid. 581–2).

These results might seem to be exactly in line with Sripada's model, but they are far from decisive. Importantly, subjects reading this extended vignette still attributed significantly lower FREE WILL and higher DEEP SELF DISCORDANCE than subjects in the NOT MANIPULATED condition. In fact, the FREE WILL scores were closer to those in the original MANIPULATED condition than to those in the NOT MANIPULATED condition, and the DEEP SELF DISCORDANCE scores *much* closer (ibid. 589). This is a little surprising if we take the added passage to straightforwardly state the conditions that people in the original MANIPULATED condition had taken manipulation to undermine (i.e. the DEEP SELF DISCORDANCE and CORRUPTED INFORMATION conditions). But it is not at all surprising if some worries about deep self discordance are driven by incompatibilist worries about free will, moral responsibility and blame, worries triggered by the manipulation. These worries would remain even if Bill were psychologically normal with respect to the explicitly stated conditions. Moreover, we do not need Sripada's hypothesis to explain why the concluding section added in the second study raised FREE WILL attributions. It is antecedently plausible that many people are generally conflicted about free will and moral responsibility, having both incompatibilist and compatibilist tendencies. We should expect a concluding section stressing compatibilist conditions to strengthen the latter tendencies.

To further test the suggestion that FREE WILL judgments explain DEEP SELF DISCORDANCE judgments, it would be helpful to try a strategy that does not rely on a fully successful manipulation of DEEP SELF DISCORDANCE judgments. One straightforward alternative is to compare causal models treating DEEP SELF DISCORDANCE as a mediator between MANIPULATION and FREE WILL with models treating FREE WILL as a mediator between MANIPULATION and DEEP SELF DISCORDANCE, to see which best fit the data.[7]

Consider next a worry about the *origins* of the motivational springs that might constitute an agents deep self. The worry is that in denying that a certain action was something an agent

[7] Sripada made no such comparison, but a search for structural equation models accounting for the covariance data published in Sripada's study (2012: 590) suggests that the best model where PDD, TRW and WDD are causally upstream from OFW, CTR and MRR is considerably better than the best model with the opposite causal order (difference in BIC scores > 6; best model: df=25, χ^2=33.3721, p=0.1220, BIC=-103.6439). This is exactly in line with Sripada's suggestion. On the other hand, the search also suggest that the best model in which there is a significant direct effect of MAN on either OFW, MRR, OR CTR (in this case, the best model overall) is *much* better than the best model without such an effect (difference in BIC scores >15), contrary to Sripada's central claim. In any case, as we shall see, there are reasons to be suspicious of the TRW and WDD measures used in Sripada's study, and so suspicious of these model comparisons. The search was performed using HBSMS, a search algorithm designed to help one find the best significant structural equation models. For further details, see section 3.

truly wanted, or wanted deep down, or something reflecting who the agent truly is, subjects might be denying that the agent is the *ultimate source* of the relevant springs of action: they came from elsewhere—in this case from Dr. Z. If this is how subjects interpret DEEP SELF DISCORDANCE statements, they might be best understood, not as tracking the violation of a compatibilist condition on free will and moral responsibility, but as tracking the violation of the seemingly incompatibilist requirement that agents be the ultimate source of their own springs of action. To see whether this is the case, we might ask subjects for their judgments about ultimate sourcehood, and determine whether such judgments affect, are affected by, or perhaps independent of, judgments of DEEP SELF DISCORDANCE.

Apart from these two general worries about DEEP SELF DISCORDANCE statements, there are specific problems with the TRULY WANT and WANTED DEEP DOWN statements, which were, statistically, the two most important statements for measuring DEEP SELF DISCORDANCE in Sripada's study (2012: 591, Table A4). The first says that "the real Bill did not truly want to kill Mrs. White—*Bill killed only because Dr. Z wanted him to*" (my italics). The immediate problem with this statement is that subjects might take the italicized clause to imply that Bill's behavior is completely caused or determined by Dr. Z, or more generally caused or explained by factors outside of Bill's control. If so, subjects would only accept the statement if they thought that an *incompatibilist* condition on free will and moral responsibility were violated. This would undermine Sripada's contention that effects of MANIPULATION on FREE WILL are mediated by subjects' sense that *compatibilist* conditions are violated. The second problematic statement, WANTED DEEP DOWN, says that "Bill is *constrained* by Dr. Z to act in a way that differs from how he himself, deep down, wants to act" (my italics) has almost exactly the same problem. Here subjects might understand 'is constrained to act so-and-so' as meaning something like 'is causally determined to act so-and-so by circumstances beyond his control'. Again, this would mean that subjects' would accept the statement only if they thought that the case violated a standard incompatibilist condition on free will and moral responsibility rather than the compatibilist condition intended by Sripada. Since most of the effect of MANIPULATION on FREE WILL in Sripada's model passed through DEEP SELF DISCORDANCE, and since the latter variable was most strongly associated with agreement with the problematic TRULY WANT and WANTED DEEP DOWN statements, a replication of Sripada's study should avoid formulations such as these.

3. A new study

To test the robustness of Sripada's results in light of the worries enumerated above and better understand why people take manipulation (of one sort) to undermine responsibility, I ran an experiment using Sripada's BILL AND DR. Z story but with somewhat different statements to track attributions of FREE WILL and DEEP SELF DISCORDANCE. Instead of the problematic CONTROL statement, I used the following desert-invoking statement as part of the FREE WILL condition:

DESERT (DES): Bill deserves to be punished for killing Mrs. White.

More importantly, however, I removed reference to Dr. Z's involvement as cause or constraint of Bill's action from Sripada's TRULY WANT and WANTED DEEP DOWN statements to lower the risk of incompatibilist interpretations, substituting the following:

TRULY WANT* (TRW*): The real Bill did not truly want to kill Mrs. White.

WANTED DEEP DOWN* (WDD*): In killing Mrs. White, Bill did not do what he wanted to do, deep down.

I also added a statement intended to capture intuitions driving source-incompatibilism about free will and moral responsibility:

OUTSIDE OF ULTIMATE CONTROL (OUC): Bill's killing of Mrs. White was ultimately explained by factors outside his control.

Like Sripada's CONTROL measure, OUTSIDE OF ULTIMATE CONTROL is spelled out in terms of 'control', which I have already suggested has both compatibilist and incompatibilist interpretations in ordinary parlance. But the notion's flexibility matters less in this context as long as factors taken to "ultimately explain" Bill's action are outside of Bill's control on either of these interpretations. Here, this is most likely the case: when subjects think of factors that might ultimately explain Bill's actions and be outside of his control, the most salient will likely be the interventions of Dr. Z, over which Bill presumably had neither kind of control. (This is not to say that the interpretation of OUTSIDE OF ULTIMATE CONTROL is unproblematic. I'll return to that issue in section 4.)

Subjects (N=361) were recruited through Amazon's Mechanical Turk (for discussion of this subject pool, see Paolacci and Chandler 2014). After reading the Bill and Dr. Z story in either the MANIPULATED or NOT MANIPULATED version, they answered the following two 'yes' / 'no' questions designed to control the accuracy of their replies:

THINK CLEARLY: Did Bill think clearly about his own desires and values before deciding to kill Mrs. White?

IMPLEMENTATION: Did Dr. Z implement his plan?

Subjects were then presented with OUTSIDE OF ULTIMATE CONTROL, FREE WILL, CORRUPTED INFORMATION, and DEEP SELF DISCORDANCE statements, in randomized order, and asked to indicate level of agreement on a 1 to 7 scale.

The questions are summarized in Table 1.

Table 1 Questions and statements

	THINK CLEARLY	Did Bill think clearly about his own desires and values before deciding to kill Mrs. White?
	IMPLEMENTATION	Did Dr. Z implement his plan?
	OUTSIDE OF ULTIMATE CONTROL (OUC)	Bill's killing of Mrs. White was ultimately explained by factors outside his control.
FREE WILL (FRW)	OWN FREE WILL (OFW)	Bill killed Mrs. White of his own free will.
	DESERT (DES)	Bill deserves to be punished for killing Mrs. White.
	MORAL RESPONSIBILITY (MRR)	Bill is morally responsible for killing Mrs. White.
CORRUPTED INFORMATION (CIN)	FALSE INFORMATION (FIN)	Bill killed Mrs. White based on false information about her, and he was deprived of any opportunity to learn the truth.
	MORAL IGNORANCE (MIG)	Bill was never taught about why certain actions are right and wrong, so he does not truly know that killing Mrs. White is wrong.
	PRACTICAL IGNORANCE (PIG)	Bill killed Mrs. White because his upbringing kept him ignorant of alternative, non-violent, ways of acting.
DEEP SELF DISCORDANCE (DSD)	PERSON DEEP DOWN (PDD)	Bill's killing of Mrs. White does not reflect the kind of person who he truly is deep down inside.
	TRULY WANT* (TRW*)	The real Bill did not truly want to kill Mrs. White.
	WANTED DEEP DOWN* (WDD*)	In killing Mrs. White, Bill did not do what he wanted to do, deep down.

Results:

301 subjects answered the control questions correctly and were included in further calculations.[8] Composite scores for FREE WILL, CORRUPTED INFORMATION and DEEP SELF DISCORDANCE were calculated for each subject by taking the means of the level of agreement

[8] N=143 / 158 for the MANIPULATED / NOT MANIPULATED conditions. 19 subjects gave the wrong answer to IMPLEMENTATION, 41 rejected THINK CLEARLY (40 of which in the NOT MANIPULATED condition). Though THINK CLEARLY is explicitly stated in the BILL AND DR. Z vignette, subjects rejecting it might have given it an interpretation on which it can reasonably be seen as false, an interpretation on which thinking clearly presupposes access to relevant information. Excluding those answering THINK CLEARLY in the negative might thus be on the conservative side, but including their answers did not meaningfully change any of the results presented here.

for the three groups of statements; each formed a reliable scale.[9] A mediation analysis was then performed using Sripada's model, i.e. treating MANIPULATION as independent variable, FREE WILL as dependent variable, and CORRUPTED INFORMATION and DEEP SELF DISCORDANCE as possible mediators. While both mediators contributed significantly to the total effect in that model, there was also a highly significant *direct* effect of MANIPULATION on FREE WILL, accounting for 43% of the total effect.[10] If CORRUPTED INFORMATION and DEEP SELF DISCORDANCE exhaust plausible compatibilist mediators, this suggests that subjects have significant incompatibilists tendencies, contrary to Sripada's conclusion.

The appearance of a considerable direct effect in this study seems to confirm worries that Sripada's results relied on the particular statements he used to measure DEEP SELF DISCORDANCE, in particular the TRULY WANT and WANTED DEEP DOWN statements, and perhaps also the CONTROL statement. But it does not yet address the further worry that agreement with DEEP SELF DISCORDANCE statements might be the result rather than the cause of intuitions of undermined free will and moral responsibility. Nor does it address the worry that agreement with DEEP SELF DISCORDANCE depends on agreement with source-incompatibilist intuitions of the sort that OUTSIDE OF ULTIMATE CONTROL might capture.

To get a better understanding of these issues, I used structural equation modeling to represent and compare a variety of possible hypothesis about the causal relations between MANIPULATION (MAN), and variables representing degrees of agreement with OUTSIDE OF ULTIMATE CONTROL (OUC) and the three statements under each of FREE WILL (FRW), CORRUPTED INFORMATION (CIN), and DEEP SELF DISCORDANCE (DSD). The models were compared with respect to their BIC value, a measure designed to balance how well a model fits with the data against the model's simplicity, in particular against how many relations of statistical dependence are explicitly represented in the model. Since our concern is with models representing causal relations, the models were all non-cyclical, and can be seen as

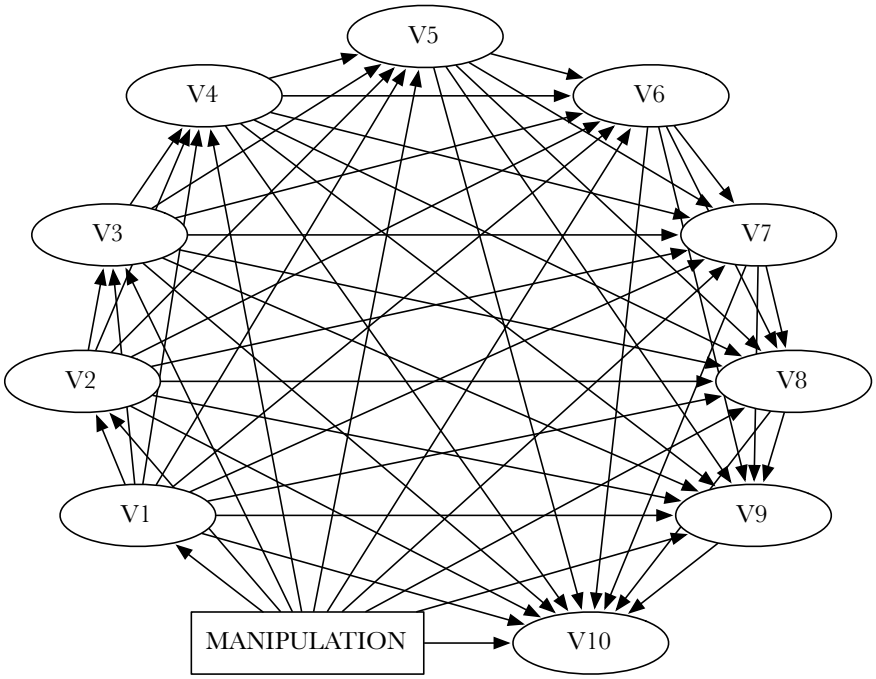
[9] Cronbach's α: .843, .828, and .841, respectively. Means for MANIPULATED / NOT MANIPULATED: FRW: 5.1/6.6 CIN: 4.9/2.7 DSD: 3.3/1.9 OUC: 4.9/2.1. Correlations ($p<.000$): MAN, OUC: -.695 MAN, CIN: -.659 MAN, FRW: .600 MAN, DSD: -.531 OUC, CIN: .723 OUC, FRW: -.683 OUC, DSD: .622 CIN, FRW: -.602 CIN, DSD: .589 FRW, DSD: -.650.

Attributions of free will and moral responsibility in the MANIPULATED condition were considerably higher than they had been in Sripada's study. The most plausible explanation of this difference is two-fold: (1) Subjects had to assess the THINK CLEARLY statement immediately after reading the vignette, leading subjects to focus on the absence of external constraints and presence of agential endorsement at the time of action, much as the appended text in Sripda's second study. (2) Subjects excluded based on the THINK CLEARLY test had almost one point lower FRW scores than others in MANIPULATION condition.

[10] With 95% confidence, the contribution of direct effect to total effect lies within a 30 to 56% interval; CI intervals calculated with Hayes's (2013) PROCESS macro for SPSS, model 4, with 10 000 percentile bootstrap intervals.

instances of the model schema in Figure 2, with each of the ten dependent variables occupying one of the variable positions (ovals), and with 0 to 55 of the possible dependence relations between variables (arrows) explicitly specified in the model. [11]

Fig. 2 Model schema



First question: Do DSD statements track intuitive conditions on moral responsibility?

Our first worry was that subjects' agreement with DSD statements plays no independent role in explaining the effects of MAN on FRW variables, being a result rather than a cause of subjects' finding Bill responsibility undermined. To see whether this might be so, I compared four kinds of models, only the last of which is in line with the expectation that DSD statements represent an independent compatibilist requirement on moral responsibility:

(i) Unconstrained models (all possibilities in the schema held open).

[11] Lower scores are better. Following Raftery 1995:139, I describe the evidence in favor of the lower-scoring model provided by a BIC difference of 2–6 as "positive", 6–10 as "strong", and >10 as "very strong". For the motivation behind BIC ("Bayesian Information Criterion"), see for example Wagenmakers 2007. Because of the vast number of possible models, the search for best-scoring models was limited to linear models and relied on GES and HBSMS search algorithms running on Tetrad 5.0.0, a freeware program for causal modeling and statistical processing; see http://www.phil.cmu.edu/projects/tetrad/. To minimize the risk of missing high-scoring models satisfying certain constraints, search parameters (Search Alpha and Beam Width) within HBSMS were systematically varied.

(ii) Models where the effect of MAN on DSD variables is entirely mediated by other variables (i.e. with no arrows leading directly from MAN to DSD variables).

(iii) Models where DSD variables have no effect on OUC or FRW variables (i.e. with no arrows leading directly or indirectly from the former to the latter variables).

(iv) Models where DSD variables have direct or indirect effect on FRW variables, and where there are no effects on DSD variables from OUC or FRW variables (i.e. with arrows leading from DSD to FRW variables, but no arrows leading from OUC or FRW to DSD variables).

The results confirm our worry. The best model of categories (i), (ii) and (iii) were all on a par, and much better than models in category (iv).[12] This provides strong evidence against the assumption that DSD variables track what people take to be an independent (compatibilist) condition on free will and moral responsibility: what they track instead seems causally downstream from what the other dependent variables track, including the FRW and OUC variables. We thus have strong reason to reject Sripada's model and his explanation of the intuitions lending force to manipulation arguments.

Second question: Does agreement with OUC influence agreement with DSD statements?

A specific part of the worry about subjects' agreement with DSD statements was that it might be influenced by their sense that Bill's action was explained by factors outside of his control (in particular the actions of Dr. Z). To test that hypothesis, I compared models where agreement with OUC influences agreement with DSD statements with models where it does not. Here, the best model of the first kind was considerably better than the best model of the second kind, providing positive evidence of OUC to DSD influence.[13]

Third question: Does agreement with OUC influence agreement with FWR statements?

A related question concerned whether manipulation negatively influenced attributions of free will, responsibility, and desert of punishment by inducing the sense that Bill's action was explained by factors outside of his control. To assess this, I compared models where OUC directly influences one or more of the FRW variables with models where it does not. The best model with such influence was much better than the best model without, providing very

[12] Scores for best model in respective category: (i) and (ii) df (degrees of freedom)=31, χ^2=39.3702, p=0.1439, BIC=-137.5502, (iii) df=30, χ^2=33.7885, p=0.2894, BIC=-137.4248, (iv) df=29, χ^2=37.3429, p=0.1377, BIC =-128.1633. BIC differences between best models and models where DSD variables play the role suggested by Sripada approach 10, corresponding to posterior odds of more than 100:1 given equal prior odds (Wagenmakers 2007: 797).

[13] Best model with OUC to DSD influence was identical to the best model overall: df=31, χ^2=39.3702, p=0.1439, BIC=-137.5502. Best model without such influence: df=29, χ^2=33.2654, p=0.2672, BIC=-132.2408. The BIC difference corresponds to a ratio in posterior odds of roughly14:1.

strong evidence that that OUC influences at least some of the FRW variables.[14] The best models overall were ones where OFW influences OUC which in turn influences MRR and DES. But the difference between these models and best model where OUC also influences OFW is too small to provide positive evidence against the latter.[15]

Fourth question: Which are most likely the largest mediators of MAN's effect on MRR?

To get a better sense of the extent to which the effect of MAN on MRR might be mediated by seemingly compatibilist and incompatibilist friendly factors, I calculated the proportion of the effect of MAN on MRR that was unmediated by the CIN and DSD variables for models that were not significantly worse than the best-scoring models. Looking specifically at (a) the best model overall (see Figure 3; numbers on arrows are coefficients for linear relations) and (b) the best model where OUC influences OFW, the lion's share of the effect of MAN on MRR were unmediated by variables designed to track salient compatibilist conditions (74% and 71%, respectively).[16] Moreover, a large proportion of the effect was mediated, in one way or other, by OUC (49% and 78%, respectively). Similar proportions were found in the other models.[17]

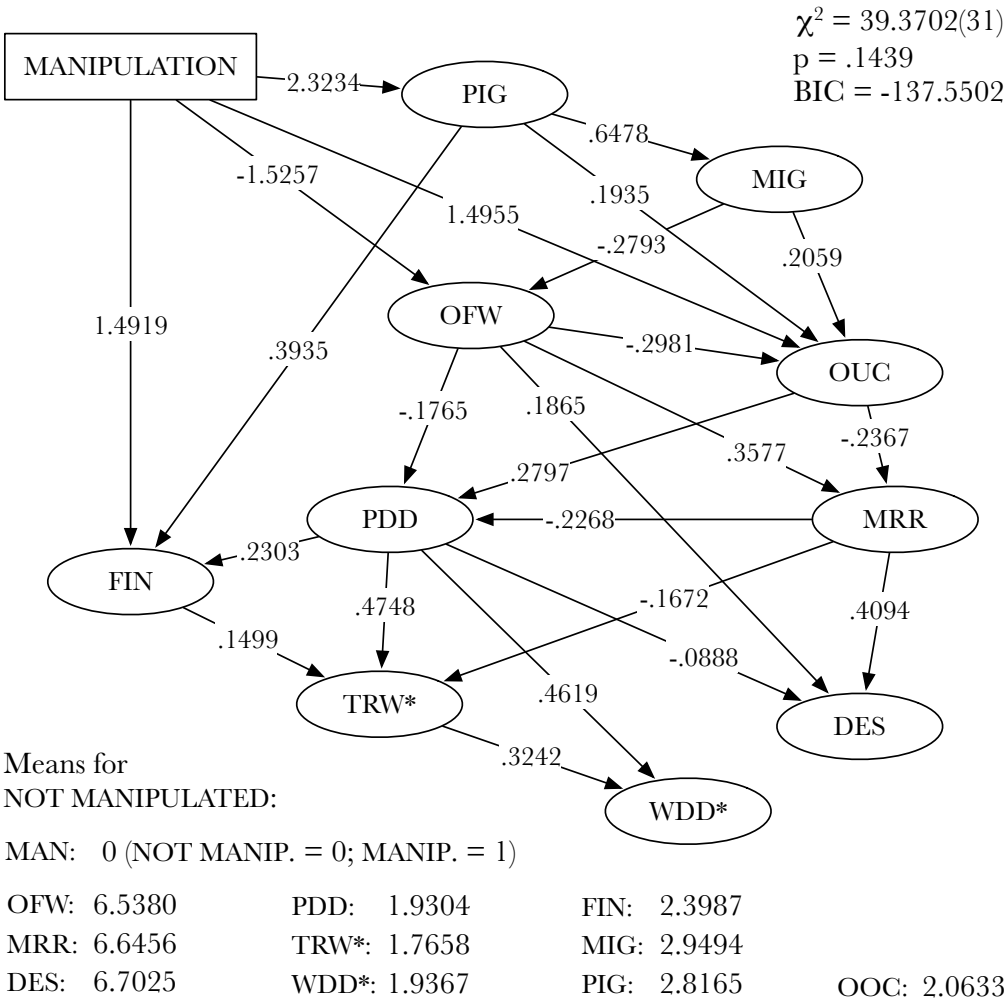
[14] Best model with direct influence from OUC to some FRW variable: df=31, χ^2=39.3702, p=0.1439, BIC=-137.5502. Best model without such influence: df=28, χ^2=38.8246, p=0.0838, BIC=-120.9745. The BIC difference corresponds to a difference in posterior odds of roughly 4000:1. Notice also that even in the best model of the latter kind, OUC still influenced FRW variables *indirectly*, via PDD.

[15] The best model of the latter kind: df=31, χ^2=40.7095, p=0.1138, BIC=-136.2109. Ratio of posterior odds between the best model overall and this model is < 2:1.

[16] With the exception of the direction of influence between OFW and OUC and resulting adjustments of coefficient strengths, the latter model is identical with model represented in Figure 3.

[17] Given the vast number of possible models and the heuristic nature of the search algorithms employed, it is possible that there are better models within the various categories investigated in this section, even though a wide range of searches were performed. To ensure the robustness of the conclusions, I also performed corresponding model comparisons operating only with MAN, OUC, and the compound variables FRW, DSD and CIN, providing a much smaller search space. The results of these comparisons were in line with the comparisons presented here: very strong evidence of a direct effect of MAN on FRW, positive evidence against a direct effect of MAN on DSD, and positive evidence of effect of OUC on DSD. Statistics for the best models: df=2, χ^2=4.2188, p=0.1213, BIC=-7.1955.

Fig. 3 Best model overall, with coefficients



4. Discussion: the manipulation argument

To avoid worries about the FREE WILL and DEEP SELF DISCORDANCE measures used in Sripada's study, I made two adjustments for the study reported here: I substituted the DESERT statement for Sripada's CONTROL statement in the FREE WILL measure, and removed possibly confounding material from the TRULY WANT and WANTED DEEP DOWN statements. The results after these adjustments were importantly different from Sripada's. As in his study, MANIPULATION had a highly significant effect on FREE WILL scores, but now a substantial part of the effect was direct, unmediated by either DEEP SELF DISCORDANCE or CORRUPTED INFORMATION. Manipulation arguments thus seem to survive the sort of test Sripada proposed. In fact, rather than supporting a compatibilist interpretation of intuitions about manipulation cases, the upshot provides prima facie support for the incompatibilist interpretation. After all, most of the effect of manipulation on attributions of free will and moral responsibility seems independent of its effects on the compatibilist conditions that Sripada thought would be most obviously undermined in BILL AND DR. Z. Instead, it was

largely direct or mediated by what seems to be a straightforwardly incompatibilist condition, OUTSIDE OF ULTIMATE CONTROL. However, let me briefly mention two ways in which one might resist this interpretation of the data:

Appeal to other compatibilist conditions: While the battery of DEEP SELF DISCORDANCE and CORRUPTED INFORMATION statements used in the study might cover the more obvious compatibilist grounds for denying that Bill was fully responsible for his action, there are several other possibilities. For example, some philosophers have proposed negative historical conditions on moral responsibility, requiring that deliberative capacities have been formed in a suitably *normal* way, or that they have not been *bypassed* in forming the agent's values or normative outlook (e.g. Haji and Cuypers 2004, 2007; Mele 1995: 166–172, 183–4). Depending on how these conditions are understood, they might be violated in BILL AND DR. Z. One might also think that the mere fact that manipulation is an *intentional intervention by another agent* is itself responsibility undermining (cf. Waller 2014). If sensitivity to such additional conditions plays a role in explaining why people take the manipulation of Bill to undermine his responsibility, and if such conditions are independently plausible conditions on responsibility, Sripada might still be right that the compatibilist has the better account of intuitions about manipulation cases.

These are speculative proposals, of course, and more research is needed to determine to what extent people are sensitive to these further conditions. More discussion is also needed to decide whether these further conditions are plausible *compatibilist* requirements for responsibility or free will. Incompatibilists or skeptics might argue that what really is responsibility undermining about intentional interventions by other agents, or by abnormal character formation or formation that bypasses agential capacities, is that it removes opportunities, or undermines our sense that the agent is the ultimate source of his action. Moreover, this sort of reply is supported by the fact that half or more of the effect of MANIPULATION on MORAL RESPONSIBILITY seemed to be mediated by what looks like judgments about an incompatibilist condition on responsibility (i.e. OUTSIDE OF ULTIMATE CONTROL). Compatibilists who want to explain (rather than explain away) the intuitive role of this condition need to argue that it is best understood under a compatibilism-friendly interpretation.[18]

Appeal to the weakness of effects: One striking aspect of both this and Sripada's study is that the effects of manipulation on FRW were modest. Between the NOT MANIPULATED and MANIPULATED conditions, the FRW score fell only 1.5 points on a 7-point scale, and

[18] Karl Persson and I (Björnsson and Persson 2012; cf. 2013) have argued that seemingly incompatibilist intuitions are best understood as resulting from pragmatic shifts in judges' explanatory frames—including their operative explanatory models—and, based on this, that compatibilist intuitions should be given more weight.

attributions of free will and moral responsibility were still well above midpoint in the MANIPULATED condition. (FRW = 5.1; in Sripada's study, the effect was larger: 2.5 for OFW and 1.7 for MRR. For discussion, see n. 9). This, one might think, is much too high to support an incompatibilist position of much significance: perhaps determinism would detract from responsibility, but not much.

Here it is worth recalling a point made in the discussion of possible problems with Sripada's study. There we noted that it is unclear whether subjects generally take Bill's actions to be *entirely* determined by Dr. Z's actions, and we now have some evidence that they did not: in the manipulation condition, the mean of OUC only reached 4.9, which is considerably closer to the midline than to complete acceptance. It might well be, then, that subjects would take responsibility and free will to be considerably more undermined if manipulation scenarios were made explicitly deterministic. More studies are clearly needed to determine whether this is the case.[19]

It is also worth noticing that the means of the variables do not tell the full story. Between the NOT MANIPULATED and MANIPULATED conditions, the proportion of subjects who "strongly agreed" with all three FRW claims went down from 51% to 7%, and the proportion who on average at least "agreed" with these claims went from 92% to 36%. These are significant shifts. They are particularly significant in light of the fact that attributions of some degree of free will and moral responsibility might be due to the existence of kinds of free will and moral responsibility that fall short of the relevant basic desert-entailing kinds.[20] Though more will have to be done to determine the interpretation of the remaining FRW attributions, the effects seem significant enough to create a prima facie problem for compatibilist accounts.

Thus far I have argued that Sripada's methodological approach is potentially helpful, but that our attempt at an improved replication of Sripada's study suggests that subjects accord

[19] A study by Feltz et al. (2013) looked at the effects of different kinds of explicitly deterministic manipulation on a composite 1–7 measure of attributions of free will. For the cases most similar to Sripada's (what Feltz et al. call *Intentional indirect manipulation* and *Culture*, attributions were at 4.28 and 4.91 respectively. These are lower measures than obtained here, in line with the suggestion that a more fully deterministic scenario would have more effect. The comparison is merely suggestive, however, as there are other differences between their cases and the one used here, and between the questions used to measure free will attributions. (Feltz et al. (2013) take the fact that deterministic manipulation undermines responsibility more than do non-manipulative deterministic scenarios to itself undermine EQUIVALENCE. For discussion, see Björnsson and Pereboom (2015).)

[20] The current study introduced a question about *deserved punishment* to strengthen the connection to basic desert, but this question too is significantly open to interpretation, as subjects could be operating with consequentialist or contractualist notions of deserved punishment rather than with basic retributivist notions.

considerable weight to incompatibilist considerations. Though the interpretation of the results is still very much up for discussion, work seems cut out mainly for compatibilists.

5. Discussion: the deep self in moral psychology

Setting aside the question of incompatibilism and the role of manipulation arguments, more should be said about the finding that judgments of deep self discordance are affected by judgments of free will, moral responsibility, and ultimate sourcehood rather than the other way around. This finding, I think, has two kinds of consequences for contemporary moral psychology.

The first consequence concerns conceptions of the deep self. Deep self accounts of responsibility standardly understand the deep self as some privileged *internal* aspect of the agent's psychology, such as higher order attitudes (Frankfurt 1971), value judgments (Watson 1975), plans (Bratman 1997), or "cares" (Shoemaker 2003). Discussions in the empirical literature seem to follow along with that idea. However, since intuitive judgments of deep self discordance of the sort canvassed here seem to be significantly affected by sourcehood worries, what matters for these judgments is not merely that the action in question reflects internal aspects of the agent. It also seems to matter whether the agent was in control of how these aspects came about. The folk psychological conception of a deep self thus seems importantly different from the philosophical conception: at least under certain circumstances, people seem to conceive of the relevant depth as not entirely internal, but as partly concerned with the *ultimate source* of the action (however ultimacy is understood). If one is primarily concerned to construct or evaluate theories of responsibility, one might want to dismiss this as in itself an insignificant deviation of a folk notion from a philosophical notion introduced specifically to distinguish different internal sources of action. But if one has assumed that the folk conception might track the philosophically important distinction (as Sripada does in his study), or hopes for one's philosophical theory to articulate an inchoate everyday notion of a deep self, one should take heed.

The second consequence is that deep self judgments might have a less fundamental explanatory role than some have thought. In our study, variations in judgments of responsibility seemed to explain variations in judgments of deep self discordance rather than the other way around. This suggests that people do not (always) decide whether an agent is responsible for something by deciding whether it reflects the agent's deep self (in the intuitive sense expressed by DSD statements): they assess the agent's responsibility on other grounds. This in turn gives us reason to be suspicious of the proposal that people take responsibility to be undermined in other contexts—in deterministic scenarios, or in light of neuroscientific explanations of action, say—because they take the agent's deep self to be disconnected from the action (for such suggestions, see Nadelhoffer et al. 2013, cf. Nahmias and Murray 2010; Murray and Nahmias 2014). But it also gives us more general reasons to be cautious about other claims about how deep self judgments affect psychological variables. For example, one

recent battery of studies of such interactions assumes rather than hypothesizes that any influence would go from deep self judgments to attributions of valuing, happiness, weakness of will, blameworthiness, and praiseworthiness, rather than the other way around (Newman, Freitas and Knobe 2014). However, as results from this study illustrate in the case of responsibility, we might know too little prior to actual testing to make confident assumptions about directions of causation. On closer scrutiny, we might thus well find that deep self judgments rarely explain, but instead are explained by, these other attributions. (This seems particularly likely in the case of attributions of blame, as these tend to go with attributions of responsibility). Given the emphatic and metaphorical nature of talk about what an agent "truly" or "really" wants, or is, "deep down," this would not be surprising.

6. Conclusion

While Sripada's argument was based on an innovative and potentially useful methodology, his study had two major weaknesses. First, two of the three statements measuring deep self discordance made reference to the constraining force of the manipulator. Second, the attempt to control for direction of causation by directly influencing judgments of deep self discordance and corrupted information was largely inconclusive.

After correcting for these weaknesses, our study yielded almost the opposite picture of Sripada's: Most of the effect of manipulation on attributions of moral responsibility was unmediated by worries about inadequate information or deep self discordance. Moreover, the remaining effect depended largely on worries that the action is ultimately explained by factors outside the agent's control, just as incompatibilists have suggested. Judging by this study, manipulation arguments might still provide a serious challenge for compatibilists. Moreover, judgments of deep self discordance were themselves explained by worries about responsibility rather than the other way around, and sensitive not only to the agent's internal psychological structure, but also, it seemed, to the source of that structure. Apparently, then, such judgments do not track any independent compatibilist condition on free will and moral responsibility, and it is unclear to what extent we should expect them to play other roles that have been proposed in the literature.

Interestingly, the problems discovered with Sripada's conclusions—problems of interpreting the prompts and figuring out the direction of causation—are similar to problems discovered with conclusions drawn in recent papers by Dylan Murray and Eddy Nahmias (Nahmias and Murray 2010; Murray and Nahmias 2014). Based on mediation analysis of the sort Sripada used, Murray and Nahmias argued that when people withhold attributions of moral responsibility and free will to agents in deterministic scenarios, they do so because they mistakenly understand determinism to imply that the agent's beliefs, desires and deliberation *have no effect* on the agent's actions. Just as Sripada found no direct effect of MAN on FRW in a model where DSD and CIN were mediators, Murray and Nahmias found no significant direct effect of deterministic scenarios on attributions of moral responsibility and free will in a model

where agreement with statements that there are no such effects was a mediator. Earlier studies have cast doubt on Murray and Nahmias' conclusion much as the results here have cast doubt on Sripada's: David Rose and Shaun Nichols (2013) present strong evidence that subjects agree with *no effect* statements about agents in deterministic scenarios because they take free will to be undermined rather than the other way around, and Gunnar Björnsson (2014) presents evidence against taking the *no effect* statements at face value, as subjects accepting such statements simultaneously tend to agree that when earlier events cause agents' actions, they do so by affecting the agents' beliefs, desires, and decisions. (Both Rose and Nichols and Björnsson also propose competing alternative explanations of why agreement with *no effect* statements would be negatively correlated with attributions of free will.) One could take the recurring problems faced by attempts to draw conclusions using mediation analysis as reason to avoid such analysis. Instead, given the possibility of comparing models involving different patterns of causal influence, our conclusion should be that analyses of causal models are useful tool for testing assumptions about the interpretation and direction of causal influence.

Acknowledgments

I thank the audience at the 4th Workshop of the Experimental Philosophy Group UK, at Bristol University, Chandra Sripada, Joshua Knobe, and two anonymous reviewers for *Philosophical Psychology* for very helpful comments. Work on this project was funded by Riksbankens Jubileumsfond.

References

Björnsson, G. and Persson, K. (2012). The Explanatory Component of Moral Responsibility. *Noûs*, 46(2), 326–54.

Björnsson, G. and Persson, K. (2013). A Unified Empirical Account of Responsibility Judgments. *Philosophy and Phenomenological Research*, 87(3). 611–39.

Björnsson, G. (2014). Incompatibilism and 'Bypassed' Agency. In A. Mele (Ed.), *Surrounding Free Will* (pp. 95–122). New York: Oxford University Press.

Björnsson, G. and Pereboom, D. (2015). Traditional and Experimental Approaches to Free Will and Moral Responsibility. In J. Sytsma and W. Buckwalter (Eds.), *Companion to Experimental Philosophy*. Blackwell.

Bratman, M. E. (1997). Responsibility and Planning. *The Journal of Ethics*, 1(1), 27-43

Cova, F. and Naar, H. (2012). Testing Sripada's Deep Self Model. *Philosophical Psychology*, 25(5), 647-59.

Feltz, A. (2013). Pereboom and premises: Asking the right questions in the experimental philosophy of free will. Consciousness and cognition, 22(1), 53–63.

Fischer, J. M. (1994). *The Metaphysics of Free Will: A Study of Control*. Oxford: Blackwell.

Fischer, J. M. (2002). Frankfurt-Type Examples and Semi-Compatibilism. In R. Kane (Ed.) *Oxford Handbook on Free Will*, (pp. 281–308). New York: Oxford University Press.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, 68(1), 5-20.

Gorin, M. (2013). What Makes an Intuition a Compatibilist Intuition? A Response to Sripada. Forthcoming in *Philosophia*, doi:10.1007/s11406-013-9446-1.

Haji, I., & Cuypers, S. E. (2004). Moral Responsibility and the Problem of Manipulation Reconsidered. *International Journal of Philosophical Studies*, 12(4), 439-464.

Haji, I., & Cuypers, S. E. (2007). Magical agents, global induction, and the internalism/externalism debate. *Australasian Journal Of Philosophy*, 85(3), 343-371.

Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*: Guilford Press.

McKenna, M. (2008). A Hard-line Reply to Pereboom's Four-Case Manipulation Argument. *Philosophy and Phenomenological Research*, 77(1), 142-159.

McKenna, M. (2013). Resisting the Manipulation Argument: A Hard-Liner Takes It on the Chin. Forthcoming in *Philosophy and Phenomenological Research*, doi:10.1111/phpr.12076.

Mele, A. (1995). *Autonomous agents: From self-control to autonomy*. Oxford: Oxford University Press.

Mele, A. R. (2005). A Critique of Pereboom's 'Four-Case Argument' for Incompatibilism. *Analysis*, 65(1), 75-80.

Mele, A. R. (2006). *Free Will and Luck*. New York: Oxford University Press.

Mele, A. (2009). Moral responsibility and agents' histories. *Philosophical Studies*, 142(2), 161-181.

Murray, D., & Nahmias, E. (2014). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research*, 88(2), 434-467.

Nadelhoffer, T., Gromet, D., Goodwin, G., Nahmias, E., Sripada, C., & Sinnott-Armstrong, W. (2013). The Mind, the Brain, and the Law. In T. Nadelhoffer (Ed.), *The Future of Punishment* (pp. 193–212): Oxford University Press.

Nahmias, E., Coates, D. J., & Kvaran, T. (2007). Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions. *Midwest Studies In Philosophy*, 31(1), 214-242.

Nahmias, Eddy and Murray, Dylan (2010) Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions. In J. Aguilar, A. Buckareff, and K. Frankish (Eds.), *New Waves in Philosophy of Action* (pp. 189–216): Palgrave Macmillan.

Nelkin, D. (2011). *Making sense of freedom and responsibility*. Oxford: Oxford University Press.

Newman, G. E., Bloom, P., & Knobe, J. (2013). Value Judgments and the True Self. Forthcoming in *Personality and Social Psychology Bulletin*, doi:10.1177/0146167213508791.

Newman, G. E., Freitas, J. D., & Knobe, J. (2014). Beliefs about the true self explain asymmetries based on moral judgment. Forthcoming in *Cognitive Science*.

Paolacci, G. and Chandler, J. (2014) Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, 23(3), 184-88.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge U.P.

Pereboom, D. (2015). A Notion of Moral Responsibility Immune to the Threat from Causal Determination. In R. Clarke, M. McKenna, and A. M. Smith (Eds.) *The Nature of Moral Responsibility: New Essays (*pp. 281–96): New York: Oxford University Press.

Pereboom, D., & Sie, M. (2013). Introduction. *Philosophical Explorations*, 16(2), 97-100.

Raftery, Adrian E. (1995). Bayesian Model Selection in Social Research. *Sociological methodology*, 25, 111-64.

Rose, D., & Nichols, S. (2013). The Lesson of Bypassing. *Review of Philosophy and Psychology*, 4(4), 599-619.

Shoemaker, D. W. (2003). Caring, Identification, and Agency. *Ethics*, 114(1), 88-118.

Sripada, C. (2012). What Makes a Manipulated Agent Unfree? *Philosophy and Phenomenological Research*, 85(3), 563-593.

Sripada, C. (2010). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, 151(2), 159-176.

Sripada, Chandra Sekhar and Konrath, Sara (2011). Telling More Than We Can Know About Intentional Action. *Mind & Language*, 26(3), 353–80.

Tognazzini, N. A. (2014). The Structure of a Manipulation Argument. *Ethics*, 124(2).

Wagenmakers, E-J. (2007). A Practical Solution to the Pervasive Problems of P Values. *Psychonomic Bulletin & Review*, 14(5), 779-804.

Waller, R. R. (2014). The Threat of Effective Intentions to Moral Responsibility in the Zygote Argument. *Philosophia*, 42(1), 209-222.

Watson, G. (1975). Free Agency. *The Journal of Philosophy*, 72(8), 205-220.

Wolf, S. (1990). *Freedom Within Reason*: Oxford University Press.