Effects of manipulation on attributions of causation, free will, and moral responsibility

Dylan Murray
University of California, Berkeley
Department of Philosophy

Tania Lombrozo
University of California, Berkeley
Department of Psychology

Abstract:

If someone brings about an outcome without intending to, is she causally and morally responsible for it? What if she acts intentionally, but as the result of manipulation by another agent? Previous research has shown that an agent's mental states can affect attributions of causal and moral responsibility *to that agent*, but little is known about what effect one agent's mental states can have on attributions to *another* agent. In Experiment 1, we replicate findings that manipulation lowers attributions of responsibility to manipulated agents. Experiments 2-7 isolate which features of manipulation drive this effect, a crucial issue for both philosophical debates about free will and attributions of responsibility in situations involving social influence more generally. Our results suggest that 'bypassing' a manipulated agent's mental states generates the greatest reduction in responsibility, and we explain our results in terms of the effects that one agent's mental states can have on the counterfactual relations between another agent and an outcome.

1. Introduction

Consider an extreme case of manipulation (Sripada, 2012; Phillips & Shaw, 2015; Woolfolk, Doris, and Darley, 2006): neuroscientists implant a device in your brain, allowing them to change your mental states (and thus your actions) at will. Using this device, the neuroscientists manipulate you to kill someone by causing you to have the desire and intention to do so. Did you act of your own free will? Are you morally responsible for the killing? Intuitively, manipulation mitigates your causal and moral responsibility, even for an action that you desired and intended to perform.

Now consider a case in which the "manipulator" is just the set of internal causal conditions inside your head that normally govern your mental states and behavior. Like the neuroscientists, these conditions cause you to perform the same action in just the same way. And suppose they do so deterministically – that is, such that given that state of the world and the laws of nature, events had to unfold exactly as they did. Did you act of your own free will? Are you morally responsible?

If your responses to these cases differ, it suggests that something about the nature of an initial causal factor (F1) – and perhaps specifically the presence or nature of that factor's mental states – can influence attributions of responsibility to a second factor (F2) in the same causal chain.[1] This raises an important question: why might factors exogenous to a particular person impact attributions of responsibility to that person?

Philosophers have discussed cases like these as they bear on the debate between compatibilists, who say that moral responsibility and free will can exist in a deterministic universe, and incompatibilists, who say that they cannot (Pereboom, 2001; Mele, 2006). Incompatibilists argue that compatibilism is susceptible to *manipulation arguments*, which begin with the intuition that being manipulated by another agent (like the neuroscientists) undermines an agent's moral responsibility and free will. If there is *no relevant difference* between manipulation and matched cases of deterministic causation (like the normal causal conditions inside your head), then determinism must undermine moral responsibility and free will as well. Generalizing, compatibilism is false.

[1] If F1 does not intend F2's behavior, but merely believes she'll bring it about by pursuing some other end (i.e., if she foresees but does not intend it), this may lessen the sense that F2 is manipulated, but nonetheless mitigate attributions of responsibility to F2 to some extent. We return to this case in the general discussion. (Throughout the paper, we assume one can foresee that one's action will have some event as an outcome – e.g., as a mere side effect – without intending that event.)

Manipulation arguments have some intuitive appeal. After all, why should it matter whether the same constraints on action come from intentional agents or natural laws and events? On the other hand, the law allows for weaker punishment in cases of entrapment (Carlon, 2007), and empirical findings suggest that manipulation is also intuitively unique. Most people assign lower moral responsibility and free will ratings to manipulated than to non-manipulated agents (Woolfolk et al., 2006; Sripada, 2012; Phillips & Shaw, 2015), but in at least some cases, a majority agree that agents in a deterministic universe can have moral responsibility and free will (Nichols & Knobe, 2007; Nahmias & Murray, 2010; Nichols, 2011; and Murray & Nahmias, 2014).

Mind-controlling evil neuroscientists are (thankfully) science fictional, but they only distill more prosaic concerns. Sunstein and Thaler (2008), for instance, recommend that the government 'nudge' people using research on heuristics and biases (e.g., increasing the number of organ donors by requiring people to check their driver's license to opt out, rather than in), and Freedman (2012) discusses how smartphone apps inspired by B. F. Skinner are making people more effective at "manipulating" their own dieting and addiction behaviors. With increased control, though, may come increased control by others. And with Freedman, we may worry about nudges becoming outright pushes. Indeed, Pereboom (2001) introduces an influential set of cases that vary from direct manipulation to cultural indoctrination to determinism, and Kane's version of the manipulation argument uses the behavioral modification of citizens in Skinner's *Walden Two* instead of evil neuroscientists (Kane, 1996). Worries in the same vein have a long history in political philosophy (Berlin, 1969; Pettit, 1997), and for their part, religious scholars have wrestled with how we could have free will and moral responsibility if the world were completely under the control of an omniscient God.

We don't hope to settle these issues here, of course – only to provide a reminder of how central manipulation is to public policy, religion, politics, the law, and everyday judgment and decision-making – and to stress the need for a more fine-grained set of questions, which is also crucial philosophically. To the extent that manipulation arguments turn on ordinary intuitions, their soundness depends on *which* features of manipulation intuitively undermine free will and responsibility (Sripada, 2012). Until we know what those features are, we don't know if determinism shares them.

Here, we aim to ascertain not only whether manipulation does impact attributions of responsibility to manipulated agents, but if so, the specific features (such as the manipulators' mental states) in virtue of which it does so. Our experiments use causal chains initiated by one factor (F1) that also involve a second factor (F2), and we vary the 'agency' of F1 and F2 (or how 'agentive' these two causal factors are), ranging from factors that are not agents at all, to humans who act intentionally and with foresight, to those who act 'fully manipulatively'. For instance, does F1's merely being a human being (rather than, e.g., a rock or robot) threaten F2's moral responsibility and free will? Is what matters instead that F1 intentionally cause F2's action? Or is interfering with or 'bypassing' F2's normal reasoning and desire-forming processes what's relevant? Without answers to these questions, we cannot adequately assess manipulation arguments, and we lack a basic understanding of interpersonal influence – i.e., of how one agent can influence attributions of responsibility to another.

1.1 Previous Work

Research on the attribution of causal and moral responsibility has focused almost exclusively on individuals in fairly isolated or simple social settings. This research has yielded valuable insights, including the robust finding that a person's mental states can influence whether that person is judged responsible for some outcome (see Young & Tsoi, 2013, for a review). People who foresee and intend the outcomes of their actions, for example, are generally judged more responsible for those outcomes than people who do not (Cushman, 2008; Lagnado & Channon, 2008, Young, Camprodon, Hauser, Pascual-Leone & Saxe, 2010; Young, Cushman, Hauser & Saxe, 2008; Pizarro & Tannenbaum, 2011). If one agent's mental states have *intra*personal effects on attributions of responsibility to that agent, we might also expect them to have *inter*personal effects on attributions of responsibility to other agents in the same causal chain. Some theoretical frameworks, for instance, include coercion as a mitigating circumstance (Shaver, 1985; Alicke 2000), and several studies have used scenarios involving coercion or manipulation to investigate other questions (e.g., Johnson et al., 1989; Young and Phillips, 2011). Work in experimental philosophy has focused primarily on determinism (e.g., Monroe and Malle, 2009; Murray and Nahmias, 2014; Nichols & Knobe, 2007), though Nahmias, Shepard, and Reuter (2014) show that people distinguish perfect (deterministic) prediction from manipulation.[2]

Murray and Nahmias (2014) suggest that most people only take determinism to threaten free will and moral responsibility when they confuse it with *bypassing*: "when one's actions are not causally dependent on one's relevant mental states and processes,

[2] Nahmias et al. (2014) show that people distinguish between actual manipulation and the mere possibility of manipulation (based on perfect prediction), supporting Frankfurt's (1969) contention that mere "*counterfactual* intervention" is not enough to threaten moral responsibility and free will.

such as one's beliefs, desires, deliberations, and decisions." Bypassing is one way an agent might exercise interpersonal influence over another – e.g., if F1 implants the desire and intention in F2 to cause an outcome through some form of mind-control.[3] Bypassing is thus a likely candidate for the key feature of manipulation that mitigates manipulees' responsibility. Murray and Nahmias (2014) don't investigate this hypothesis, but it is consistent with what relevant work has been done: Feltz (2013), Sripada (2012), Phillips & Shaw (2015), and Woolfolk, Doris, and Darley (2006).

Feltz (2013) directly investigates Pereboom's (2001) Four Case manipulation argument, and finds significantly different attributions of free will, moral responsibility, and blame between all four cases – direct manipulation (implantation of desires), indirect manipulation (genetic programming), cultural training, and determinism. Establishing such differences does not explain what drives them, as we attempt to here, though Feltz does report that being (directly and indirectly) "manipulated" by a brain tumor mitigates responsibility attributions significantly less than manipulation by an intentional agent.[4]

Sripada (2012) also finds that an agent (F2) is assigned less free will and moral responsibility for killing a woman when F2 is manipulated to do so by an evil scientist (F1) than when the scientist never implements his plan. Sripada suggests that the mitigating effect of manipulation is mediated by judgments about whether F2's action was in accordance with his 'deep self' (i.e., reflects the kind of person he truly is) and

[3] In order to make an implanted desire (or intention) behaviorally efficacious, the manipulator needs to make it considerably stronger than, or somehow causally isolated from, any influence of the manipulee's *other* desires (or intentions) which, if stronger, might overpower it. Thus, manipulation doesn't involve bypassing of the *implanted* desire or mental state, but in order to actually affect behavior, it will typically require bypassing the agent's other, "normal" mental states. Bypassing can be brought about in other ways – e.g., by non-human factors – but bypassing by another human agent is a type of interpersonal influence.

[4] Feltz (2013) finds that the only case in which the manipulee is intuitively *not* responsible (as indicated by an average score below the mid-point on a composite of the free will, moral responsibility, and blame questions) is when he is directly intentionally manipulated.

whether F1's indoctrination distorted the information available to F2. Similarly, Phillips & Shaw (2015) find that a group of workers receive lower blame ratings for attacking a village when its government intentionally manipulates them to do so, compared to cases in which the government's influence is unintentional (or deviant). Also consistent with these results, Woolfolk et al. (2006) find that one hostage's responsibility for killing another decreases the more effectively his alternatives are restricted by a group of hijackers – that is, to the extent that the outcome was under the control of the hijackers', rather than the hostage's, intentions.

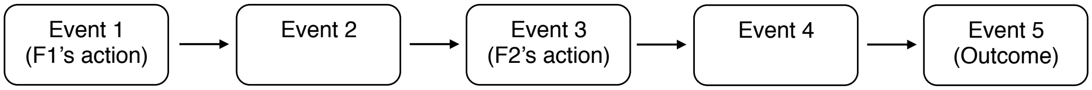
1.2 Experimental hypotheses and overview

Previous research provides some evidence suggesting that manipulation threatens the perceived responsibility of "manipulees," and that this effect depends on the manipulator's mental states, in particular her intentions. However, this work leaves open two important questions. First, *which* specific intention(s) are threatening? Second, what – if anything – is *unique* about the interpersonal effect of manipulation?

Our strategy in addressing these questions is somewhat exploratory and bottom-up: our experiments isolate important aspects of agentiveness, allowing us to see which impact responsibility. We begin by testing for interpersonal effects in canonical cases of manipulation (Experiment 1), and subsequent experiments isolate the contributions of the following six aspects of agentiveness: being an agent (Experiment 2), acting intentionally (Experiment 3), foreseeing the effects of one's actions (Experiment 4), intending another agent's action (Experiment 5), intending for this agent's action to produce the outcome (Experiment 6), and bypassing the other agent's mental states to do so (Experiment 7).

The seven experiments we present use variations on the same six vignettes, each of which describe causal chains comprised of five events in which each event directly causes the next (see Fig. 1): F1's immediate action (first event) has a proximal effect (second event) that brings about F2's immediate action (third event), the proximal effect of which (fourth event) then causes the outcome (fifth event), the death of a third party.

Figure 1: Structure of causal chains used in all experiments.



Each experiment independently varies whether the causal factors involved in the chain are more or less 'agentive' (F1 status: + or -, F2 status: + or -), spanning the range between canonical cases of *full manipulation* (which include bypassing, like the scenarios involving evil neuroscientists) and cases of 'purely deterministic', completely non-agentive causal influences (like natural laws and past events involving non-humans). Crossed with our 6 vignettes, this 2x2 design generates 24 scenarios for each experiment. The full set of variations across and within experiments is shown in Table 1 (which we recommend consulting in conjunction with the concrete example in Exp. 1; see Supplementary Material A for a full summary of all vignettes and questions).

To foreshadow our results, we find that several aspects of agentiveness drive intrapersonal and interpersonal effects, including intentionally influencing another agent or the outcome in any way. However, some of these may not be effects of manipulation, *per se*, which further analyses in the General Discussion suggest derive from bypassing.

Table 1: Overview of all seven experiments with the status of F1 and F2 indicated for each condition. '+' refers to the more agentive condition; '-' to the less agentive.

Experiment			Event 1	Event 2	Event 3	Event 4	Event 5
			F1 action	**F1 effect**	**F2 action**	**F2 effect**	**Outcome**
Example			Turns on device	Sends subliminal signal	Causes to hate	Shoves into pitchfork	Person dies
Exp. 1	**F1**	-	Non-agent	Not foreseen, not intended			
		+	Intentional	Foreseen, intended by bypassing			
	F2	-			Non-agent	Not foreseen, not intended	
		+			Intentional	Foreseen, intended	
Exp. 2	**F1**	-	Non-agent	Not foreseen, not intended			
		+	Accidental				
	F2	-			Non-agent	Not foreseen, not intended	
		+			Accidental		
Exp. 3	**F1**	-	Accidental	Not foreseen, not intended			
		+	Intentional				
	F2	-			Accidental	Not foreseen, not intended	
		+			Intentional		
Exp. 4	**F1**	-	Intentional	Not foreseen, not intended		Not foreseen, not intended	
		+		Foreseen, not intended			
	F2	-			Intentional	Not foreseen, not intended	
		+				Foreseen, not intended	
Exp. 5	**F1**	-	Intentional	Foreseen, not intended		Not foreseen, not intended	
		+		Foreseen, intended			
	F2	-			Intentional	Foreseen, not intended	
		+				Foreseen, intended	
Exp. 6	**F1**	-	Intentional	Foreseen, intended		Foreseen, not intended	
		+				Foreseen, intended by altering environment	
	F2	-			Intentional	Foreseen, not intended	
		+				Foreseen, intended	
Exp. 7	**F1**	-	Intentional	Foreseen, intended		Foreseen, intended by altering environment	
		+				Foreseen, intended by bypassing	
	F2	-			Intentional	Foreseen, not intended	
		+				Foreseen, intended	

2. Experiments

2.1 Experiment 1: Full Manipulation

Experiment 1 investigates whether 'full manipulation' by F1, in which every aspect of agentiveness is present (including bypassing), lowers attributions of responsibility to F2 compared to cases in which F1's influence is 'purely deterministic' (completely non-agentive and non-manipulative). This allows us to determine whether *at least one* feature of manipulation threatens responsibility. Experiments 2-7 then isolate *which* feature(s).

Methods

Participants. Four-hundred-and-eighty participants (mean age = 27, 66.25% female) were recruited from Amazon Mechanical Turk, a platform where participants can be recruited and compensated for performing online tasks. Participants were restricted to those with IP addresses in the United States and an approval rating of 95% or higher. An additional 144 participants were excluded for one or more of the following reasons: failing either of two comprehension questions (described below), leaving one or more items blank, or having a repeat IP address. The methods and analyses for Experiments 2-7 were identical to those in Experiment 1, save for variations in agentiveness.[5]

Materials. Materials consisted of six distinct vignettes created to appear in one of four versions. First, an initial causal factor in a causal chain, F1, is either a 'fully manipulative' human being who intentionally brings about the outcome by causing F2's action through bypassing (F1+) or a non-human (animal, inanimate object, or robot) that

[5] In Exp. 1, 15 subjects were excluded due to experimenter error. The majority of participants in all experiments were excluded because they failed a comprehension question or had a repeat IP address. Other work using Mechanical Turk has reported comparable exclusion rates (e.g., Downs et al., 2010). An ANOVA on exclusion rates across all experiments with study (7), vignette (6), F1 status (2: F1+, F1-), and F2-status (2: F2+, F2-) as between-subjects factors revealed only a significant main effect of study (the number of subjects excluded is reported in each experiment).

causes F2's action non-intentionally (F1-). A second factor in the chain, F2, is either a human being who intentionally brings about the outcome (F2+) or a non-human that causes the outcome non-intentionally (F2-). These two variations were crossed to create four versions of each vignette.

Below is the F1+ / F2+ version of the 'Manhattan' vignette, with a synopsis of the other conditions in Table 2 (see Supplementary Material A for additional stimuli):

> Cedric is having breakfast at an outdoor bistro in the middle of Manhattan. Looking up from his morning papers, he can see the skyscraper behind the businessman and his dog who are sitting across from him. Just then, at the very top of the skyscraper, a tourist with excellent aim and foresight intentionally drops a pill containing a drug as part of a complex plan to kill Cedric, since he doesn't like New Yorkers. The pill rolls over the edge of the skyscraper and falls all 47 floors right into a large cup of coffee at their table. Because the drug in the pill makes him desperately want to kill someone, the businessman sitting across from Cedric intentionally jumps up from the table, which knocks it over and sends Cedric straight backwards, just as the tourist foresaw. Cedric hits his head against the asphalt behind him – hard – and dies before help can arrive.

This description was followed by two additional paragraphs of text that reinforced the causal and intentional structure of the vignettes, repeating which events F1 and F2 respectively did and did not intend (and foresee), and the relevant counterfactual relations between them (e.g., that the tourist would not have caused Cedric's death if his action had not led to the businessman's action).

Table 2: Sample stimuli from one vignette ('Manhattan') in Experiment 1. The causal chains were presented to participants in the form of a narrative text.

Condition	F1 action	→ F1 proximal effect	→ F2 action	→ F2 proximal effect	→ Outcome
F1+, F2+	A tourist drops a drug-laden pill into a cup of coffee in order to kill Cedric	The drug makes the businessman who drinks it want to kill someone	The businessman jumps up and knocks a table over in order to kill Cedric	The table knocks Cedric backward onto his head	Cedric dies
F1+, F2-	A tourist drops a drug-laden pill into a cup of coffee in order to kill Cedric	The drug makes the dog who drinks it automatically jump up	The dog's jumping up accidentally knocks a table over	The table knocks Cedric backward onto his head	Cedric dies
F1-, F2+	A pigeon accidentally drops a nut into a cup of coffee	The coffee splashes onto a businessman, who'd rather kill someone than be scalded	The businessman jumps up and knocks a table over in order to kill Cedric	The table knocks Cedric backward onto his head	Cedric dies
F1-, F2-	A pigeon accidentally drops a nut into a cup of coffee	The coffee splashes onto a dog, whose reflexes make him automatically jump up	The dog's jumping up accidentally knocks a table over	The table knocks Cedric backward onto his head	Cedric dies

Dependent measures included the following questions about causation, free will, moral responsibility, and blame. Participants were asked to indicate their agreement with a series of statements on a 1-7 scale, with points labeled 'strongly disagree', 'disagree', 'somewhat disagree', 'neither agree nor disagree', 'somewhat agree', 'agree', and 'strongly agree'. The statements for the F1+ / F2+ version of the 'Manhattan' vignette were as follows (with variations for the other three conditions in brackets):

Moral responsibility:

The tourist [pigeon] is morally responsible for Cedric's death.

The businessman [dog] is morally responsible for Cedric's death.

Free will:

The tourist [pigeon] exercised free will in bringing about Cedric's death.

The businessman [dog] exercised free will in bringing about Cedric's death.

Blame:

The tourist [pigeon] deserves to be blamed for Cedric's death.

The businessman [dog] deserves to be blamed for Cedric's death.

Causation:

The tourist [pigeon] caused Cedric's death.

The businessman [dog] caused Cedric's death.

The tourist [pigeon] caused the businessman [dog] to jump up.

The businessman [dog] caused the tourist [pigeon] to drop the pill [nut].

Finally, each vignette had two corresponding comprehension statements, one of which was true and the other false. For the 'Manhattan' vignette, the statements were:

Comprehension:

The pill [nut] fell into a large cup of coffee.

Cedric was having breakfast in the French countryside.

Procedure. Participants were randomly assigned to a single version of a single vignette. With the vignette still visible, participants responded to the set of causation questions in a random order and the moral responsibility, free will, and blame questions in a random order, with the order of these two sets also randomized. Participants then received the two comprehension questions with the vignette removed. Finally, participants indicated their age and sex and received debriefing information.

Results

Preliminaries. Table 3 presents means and standard deviations for all measures in all experiments.

Table 3: Means and standard deviations for all experimental conditions. Ratings given on a 7-point scale, from 'strongly disagree' (1) to 'strongly agree' (7).

	Condition	F1+, F2+		F1+, F2-		F1-, F2+		F1-, F2-	
	Factor	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**
Exp. 1	Responsibility (composite)	6.33 (.86)	3.52 (1.46)	6.59 (.70)	2.06 (1.06)	2.60 (1.42)	5.74 (1.29)	3.34 (1.34)	2.61 (1.28)
	Moral Responsibility	6.35 (1.11)	3.55 (1.81)	6.71 (.75)	1.59 (1.19)	2.13 (1.58)	5.74 (1.58)	2.66 (1.85)	2.18 (1.61)
	Free Will	6.13 (1.43)	2.64 (1.55)	6.43 (1.29)	1.65 (1.23)	2.52 (1.77)	5.62 (1.57)	2.38 (1.67)	2.20 (1.60)
	Deserves Blame	6.43 (.92)	3.38 (1.87)	6.62 (1.01)	1.80 (1.33)	2.70 (1.89)	5.55 (1.67)	3.62 (2.07)	2.42 (1.75)
	Causation	6.42 (1.07)	4.52 (1.80)	6.60 (.79)	3.21 (1.94)	3.07 (1.93)	6.03 (1.32)	4.69 (1.83)	3.64 (1.92)
	F1 caused F2 to [outcome]	6.10 (1.30)		6.45 (1.08)		4.00 (2.10)		5.31 (1.60)	
	F2 caused F1 to [F1's action]	1.80 (1.39)		1.87 (1.55)		2.25 (1.72)		1.71 (1.29)	
Exp. 2	Responsibility (composite)	3.45 (1.49)	2.75 (1.30)	3.55 (1.53)	2.66 (1.25)	3.70 (1.30)	2.77 (1.39)	3.18 (1.48)	2.62 (1.17)
	Moral Responsibility	3.59 (2.02)	2.68 (1.67)	3.57 (1.91)	1.98 (1.47)	3.10 (1.92)	2.58 (1.71)	2.63 (1.82)	2.18 (1.54)
	Free Will	2.78 (1.88)	2.24 (1.52)	2.82 (1.71)	2.08 (1.43)	2.46 (1.71)	2.48 (1.73)	2.22 (1.64)	1.85 (1.23)
	Deserves Blame	3.18 (1.82)	2.44 (1.61)	3.46 (1.96)	2.48 (1.75)	4.31 (1.90)	2.39 (1.65)	3.33 (2.06)	2.53 (1.69)
	Causation	4.23 (1.80)	3.66 (1.90)	4.37 (1.90)	4.12 (2.09)	4.94 (1.69)	3.64 (2.00)	4.56 (1.81)	3.94 (1.85)
	F1 caused F2 to [outcome]	4.60 (1.84)		5.04 (1.70)		5.68 (1.36)		5.30 (1.70)	
	F2 caused F1 to [F1's action]	1.59 (1.10)		1.61 (1.18)		1.65 (1.31)		1.47 (1.07)	
Exp. 3	Responsibility (composite)	3.67 (1.51)	3.63 (1.61)	3.85 (1.59)	2.73 (1.31)	2.99 (1.39)	3.96 (1.55)	2.91 (1.33)	2.86 (1.39)
	Moral Responsibility	3.42 (1.90)	3.41 (1.94)	3.68 (2.02)	2.31 (1.57)	2.76 (1.68)	3.68 (2.00)	2.81 (1.73)	2.69 (1.76)
	Free Will	3.96 (1.82)	3.81 (1.89)	3.89 (1.94)	2.71 (1.83)	3.08 (1.86)	4.18 (1.80)	2.65 (1.75)	2.41 (1.71)
	Deserves Blame	3.21 (1.86)	3.15 (1.93)	3.48 (2.01)	2.43 (1.75)	2.63 (1.62)	3.59 (1.92)	2.53 (1.66)	2.54 (1.70)
	Causation	4.10 (1.88)	4.15 (1.86)	4.35 (1.92)	3.49 (1.87)	3.48 (1.82)	4.39 (1.87)	3.65 (1.88)	3.78 (2.00)
	F1 caused F2 to [outcome]	4.33 (1.87)		4.65 (2.01)		4.31 (1.88)		4.31 (1.81)	
	F2 caused F1 to [F1's action]	1.80 (1.40)		1.34 (.67)		1.49 (1.00)		1.39 (.79)	
Exp. 4	Responsibility (composite)	4.12 (1.54)	4.20 (1.54)	3.89 (1.55)	4.06 (1.59)	3.63 (1.68)	3.92 (1.69)	3.45 (1.48)	3.92 (1.61)
	Moral Responsibility	3.98 (1.98)	4.03 (1.96)	3.89 (1.91)	4.02 (2.00)	3.49 (1.93)	3.73 (1.97)	3.28 (1.97)	3.60 (2.04)
	Free Will	4.36 (1.81)	4.29 (1.86)	4.20 (1.75)	4.24 (1.74)	3.98 (1.92)	4.07 (1.90)	3.91 (1.93)	4.18 (1.87)
	Deserves Blame	3.73 (1.91)	3.74 (1.94)	3.43 (1.93)	3.52 (1.97)	3.27 (1.93)	3.48 (2.00)	2.95 (1.83)	3.38 (1.90)
	Causation	4.43 (1.91)	4.76 (1.80)	4.04 (1.98)	4.48 (1.97)	3.76 (1.98)	4.41 (1.94)	3.65 (1.86)	4.52 (2.01)
	F1 caused F2 to [outcome]	4.88 (1.80)		4.37 (1.97)		4.53 (2.00)		4.11 (1.99)	
	F2 caused F1 to [F1's action]	1.68 (1.33)		1.41 (.94)		1.58 (1.14)		1.43 (.85)	
Exp. 5	Responsibility (composite)	4.95 (1.48)	5.47 (1.63)	5.43 (1.27)	3.59 (1.75)	3.86 (1.65)	5.83 (1.39)	4.45 (1.60)	4.02 (1.71)
	Moral Responsibility	5.13 (1.73)	5.54 (1.87)	5.48 (1.46)	3.41 (1.95)	3.65 (2.00)	5.74 (1.79)	4.42 (1.99)	3.89 (2.05)
	Free Will	4.93 (1.83)	5.34 (1.80)	5.17 (1.67)	3.71 (1.94)	4.13 (1.93)	5.81 (1.60)	4.54 (1.85)	4.22 (1.96)
	Deserves Blame	4.78 (1.85)	5.31 (1.86)	5.45 (1.58)	3.26 (1.91)	3.63 (1.92)	5.72 (1.70)	4.15 (1.96)	3.57 (2.06)
	Causation	4.95 (1.66)	5.70 (1.71)	5.63 (1.41)	3.98 (1.97)	4.05 (2.04)	6.07 (1.44)	4.68 (1.67)	4.42 (1.93)
	F1 caused F2 to [outcome]	5.27 (1.60)		5.84 (1.49)		4.77 (2.02)		5.23 (1.75)	
	F2 caused F1 to [F1's action]	2.03 (1.48)		1.60 (.98)		1.63 (1.23)		1.59 (.97)	
Exp. 6	Responsibility (composite)	6.11 (1.16)	4.87 (1.74)	6.30 (.94)	3.74 (1.81)	5.13 (1.41)	5.44 (1.51)	5.74 (1.02)	3.59 (1.80)
	Moral Responsibility	6.25 (1.28)	4.88 (2.06)	6.47 (1.01)	3.33 (2.18)	5.33 (1.64)	5.45 (1.78)	6.03 (1.08)	3.46 (2.04)
	Free Will	6.03 (1.45)	4.72 (2.12)	6.37 (1.08)	4.09 (2.06)	4.81 (1.92)	5.43 (1.79)	5.43 (1.53)	3.71 (1.96)
	Deserves Blame	6.08 (1.29)	4.68 (2.01)	6.18 (1.27)	3.37 (2.09)	5.08 (1.66)	5.23 (1.76)	5.83 (1.19)	3.37 (2.09)
	Causation	6.08 (1.38)	5.23 (1.69)	6.21 (1.22)	4.18 (2.09)	5.30 (1.67)	5.63 (1.59)	5.68 (1.25)	3.84 (2.09)
	F1 caused F2 to [outcome]	5.45 (1.69)		5.43 (1.78)		5.18 (1.97)		5.73 (1.58)	
	F2 caused F1 to [F1's action]	2.12 (1.40)		1.68 (1.22)		2.14 (1.39)		2.07 (1.47)	
Exp. 7	Responsibility (composite)	6.50 (.72)	3.63 (1.63)	6.50 (.71)	3.22 (1.55)	5.90 (1.07)	5.42 (1.43)	6.10 (1.20)	3.56 (1.61)
	Moral Responsibility	6.60 (.69)	3.58 (1.89)	6.60 (.75)	3.07 (1.84)	6.01 (1.33)	5.48 (1.68)	6.30 (1.29)	3.22 (1.87)
	Free Will	6.43 (1.08)	2.78 (1.89)	6.34 (1.16)	2.81 (1.81)	5.87 (1.47)	5.23 (1.77)	5.88 (1.72)	3.56 (1.84)
	Deserves Blame	6.47 (1.05)	3.61 (1.92)	6.51 (.84)	2.93 (1.81)	5.87 (1.32)	5.30 (1.76)	6.14 (1.43)	3.31 (1.91)
	Causation	6.52 (.78)	4.57 (1.85)	6.55 (.73)	4.06 (1.86)	5.87 (1.37)	5.66 (1.55)	6.08 (1.37)	4.17 (1.89)
	F1 caused F2 to [outcome]	6.15 (1.05)		6.34 (.97)		5.40 (1.60)		5.93 (1.59)	
	F2 caused F1 to [F1's action]	1.78 (1.24)		1.67 (1.20)		2.53 (1.83)		1.84 (1.40)	

Due to the large number of measures, we averaged the moral responsibility, free will, blame, and causation judgments for F1 and F2 to create a single composite *responsibility measure* for each (including only the first two ‘Causation’ questions above). This decision was supported by a reliability analysis, which revealed that ratings for these judgments had excellent internal consistency (Cronbach's $\alpha = .903$).[6]

In each experiment, we analyze the composite responsibility measure for F2 as the dependent variable in a 2 (F1 status) x 2 (F2 status) x 6 (vignette) between-subjects ANOVA. This analysis allows us to address two key questions. First, to investigate whether and when "manipulation" is mitigating, we consider the effects of F1's status (more or less agentive) on attributions of responsibility to F2. This is the key *inter*personal effect of interest. Second, to investigate more standard *intra*personal effects, we consider how F2's agentiveness influences attributions of responsibility to F2. We also report main effects of vignette and its interactions with other variables, but postpone their consideration until the General Discussion.

Our experimental design also allows us to ask parallel questions with F1 as the dependent variable – that is, to test for the effects of F2 status and F1 status on *F1* responsibility ratings. Such analyses are important in establishing which effects of F2 on F1 are unique to manipulation, an issue to which we return in the General Discussion.

Effects of agency on ratings for F2. To analyze effects of agentiveness on F2 responsibility ratings, we performed an ANOVA with F1 agentive status (2: F1+, F1-), F2 status (2: F2+, F2-), and vignette (6) as between-subjects factors.

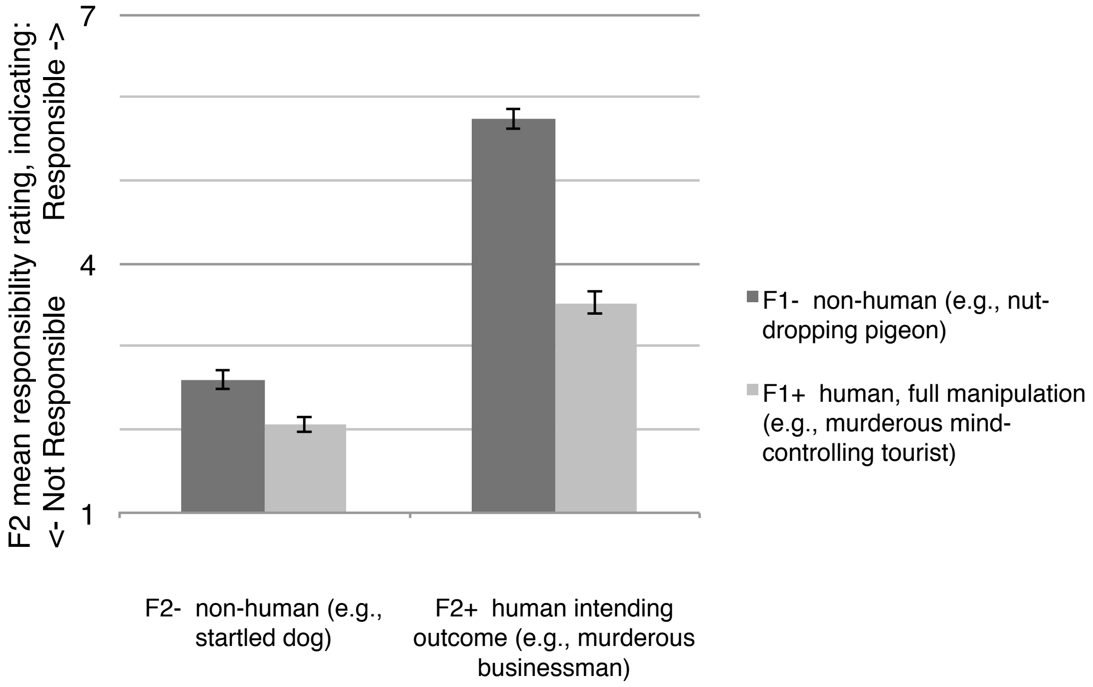
This analysis revealed significant main effects of F2 status, $F(1, 476) = 427.69$, $\eta_p^2 = .48$, $p < .001$, with higher ratings in F2+ ($M = 4.63$, $SD = 1.77$) than F2- ($M = 2.34$, $SD = 1.20$), and of F1 status, $F(1, 476) = 155.00$, $\eta_p^2 = .25$, $p < .001$, with lower ratings

[6] For F1 ratings, Cronbach's $\alpha = .915$. Reliability analyses for each experiment on the moral responsibility, free will, blame, and causation ratings revealed acceptable to excellent internal consistency: α's of .74-.92 for F2 ratings, and of .79-.88 for F1 ratings. We therefore adopted the same composite responsibility measure for analyses in all experiments.

of F2 in F1+ (*M* = 2.79, *SD* = 1.47) than F1- (*M* = 4.17, *SD* = 2.02). For example, participants tended to agree that F2 was more responsible when F2 was *a murderous human* than *a startled dog*, and less responsible when manipulated by the *mind-controlling tourist* than when prompted to act by the *nut-dropping pigeon*.

There was also a significant interaction between F1 status and F2 status, *F*(1, 476) = 56.65, η_p^2 = .11, *p* < .001: the status of F1 had a greater impact on F2 ratings in F2+ (a difference of 2.21 points) than F2- (a difference of .55 points). This could reflect a floor effect: F2 ratings were well below the mid-point in F2-. Finally, there was a significant main effect of vignette, *F*(1, 476) = 5.93, η_p^2 = .06, *p* < .001, and a significant interaction between vignette and F2 status, *F*(1, 476) = 5.85, η_p^2 = .06, *p* < .001.

Figure 2: Effect of F1 status (+, -) on mean F2 responsibility ratings in Experiment 1. There were significant main effects of F1 status and F2 status, as well as a significant interaction between F1 status and F2 status (see text). Error bars represent one SEM in each direction.



Discussion

Experiment 1 asked two questions: First, is F2's perceived responsibility sensitive to the difference between being 'fully manipulated' versus being caused by completely non-manipulative, non-human factors? Yes: F2 received lower responsibility ratings when F1 was more agentive.

Second, is F2's perceived responsibility sensitive to whether F2 is a human who intentionally brings about the outcome versus a non-human causal factor? Yes: F2 received higher responsibility ratings when F2 was more agentive.

In short, we succeeded in finding evidence that manipulation mitigates responsibility (an *inter*personal effect), and that a factor's agentive status affects its own responsibility (an *intra*personal effect). Both findings are consistent with prior research and lay the groundwork for our subsequent experiments.

2.2 Experiment 2: Being an agent

Experiment 1 compared cases of 'full manipulation' against situations in which F1 and F2 were not even human agents. In subsequent experiments, we aim to identify *which* features of agentiveness contribute to its effect on responsibility ratings. In particular, Experiment 2 tests the hypothesis that simply being an agent – having a bare intra or inter*person*al influence – impacts responsibility attributions. Experiment 2 builds on Experiment 1, using the same less agentive (-) conditions (e.g., where F1 is a nut-dropping pigeon, and F2 a startled dog). The more agentive (+) conditions of Experiment 2 add in that these causal factors are humans, but who act completely accidentally (e.g., F1 is an accidentally-nut-dropping tourist, and F2 a startled businessman).

Methods

The methods and analyses for Experiments 2-7 were identical to those from Experiment 1, except as noted.

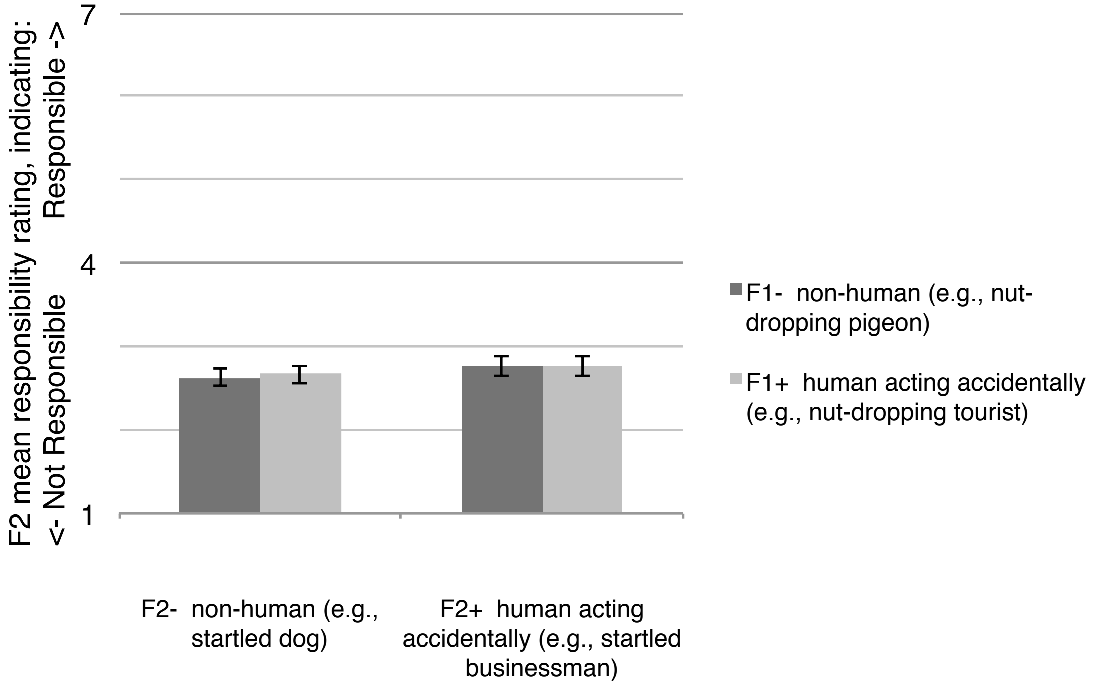
Participants. Four-hundred-eighty participants were recruited from Amazon Mechanical Turk as in Experiment 1 (mean age = 31, 52.40% female, 127 excluded).

Materials. The four conditions of the six vignettes had the following characteristics: F1 is either a human (F1+) or a non-human (F1-), and F2 is either a human (F2+) or a non-human (F2-). Both causal factors behave entirely non-intentionally and without foresight in all conditions.

Results

Effects of agency on ratings for F2. A 2x2x6 ANOVA on F2 responsibility ratings with F1 status, F2 status, and vignette as between-subjects factors revealed only a significant main effect of vignette, $F(1, 476) = 6.75$, $\eta_p^2 = .07$, $p < .001$, with no significant effects of F1 status ($p = .919$), F2 status ($p = .291$), nor an interaction between F1 status and F2 status ($p = .790$).

Figure 3: Effect of F1 status (+, -) on mean F2 responsibility ratings in Experiment 2. There were no significant main effects of F1 status or F2 status, nor any interaction between them (see text). Error bars represent one SEM in each direction.



Discussion

Experiment 2 asked two questions: First, is F2's perceived responsibility sensitive to the difference between being caused to bring about an outcome by a human being versus a non-human? No: At least in the cases here, F2's responsibility ratings did not differ significantly depending on F1's agentive status. These findings rule out one candidate aspect of agentiveness that might drive interpersonal effect(s) of manipulation: the mere fact of being an agent.

Second, is F2's perceived responsibility sensitive to whether F2 is a human being versus a non-human? No: F2's responsibility ratings did not differ significantly depending on F2's agentive status.

2.3 Experiment 3: Acting Intentionally

Experiment 3 tests the hypothesis that intending one's immediate action, without intending or foreseeing any of its effects, has an effect on responsibility ratings. Experiment 3 builds on Experiment 2, with the more agentive conditions of Experiment 2 (e.g., where F1 is a tourist who drops a nut accidentally, and F2 is a startled businessman who accidentally jumps up) becoming the less agentive conditions of Experiment 3. The more agentive conditions of Experiment 3 add in that the agents intend their immediate actions, but without foreseeing or intending any of their consequences (e.g., the tourist drops the nut intentionally because it's a bad nut, which only happens to fall in the coffee, and the businessman jumps up intentionally only to avoid being scalded).

Methods

Participants. Four-hundred-eighty participants were recruited from Amazon Mechanical Turk as in Experiment 1 (mean age = 30, 53.96% female, 174 excluded).

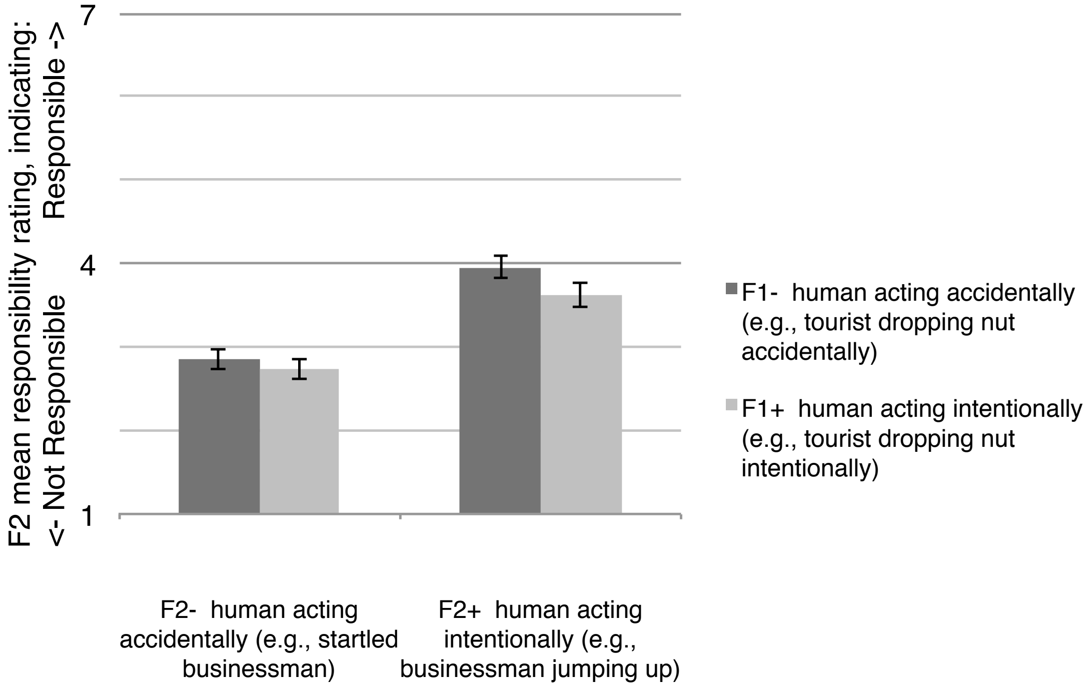
Materials. The four conditions of the six vignettes had the following characteristics: F1 is a human whose action (event 1; see Fig. 1) is intentional (F1+) or non-intentional (F1-), and F2 is a human whose action (event 3) is intentional (F2+) or non-intentional (F2-). Neither F1 nor F2 foresees nor intends any of the other events in the causal chain in any condition.

Results

Effects of agency on ratings for F2. A 2x2x6 ANOVA on F2 responsibility ratings with F1 status, F2 status, and vignette as between-subjects factors revealed a significant main effect of F2 status, $F(1, 476) = 70.78$, $\eta_p^2 = .13$, $p < .001$, with higher ratings in F2+ ($M = 3.79$, $SD = 1.59$) than F2- ($M = 2.79$, $SD = 1.35$). For example, the businessman was held more responsible when he jumped up intentionally than when he did so accidentally, despite never intending or foreseeing that this would lead to Cedric's death. There was no main effect of F1 status ($p = .057$), nor an interaction between F1 status and F2 status ($p = .381$).[7] There was a significant main effect of vignette, $F(1, 476) = 24.98$, $\eta_p^2 = .22$, $p < .001$, and a significant interaction between vignette and F2 status, $F(1, 476) = 2.95$, $\eta_p^2 = .03$, $p = .012$.

[7] The main effect of F1 status is marginal and in the expected direction, so would be significant with a one-tailed test. Given our large sample size, however, our statistical power is relatively high; if this is a real effect, it's likely small.

Figure 4: Effect of F1 status (+, -) on mean F2 responsibility ratings in Experiment 3. There was a significant main effect of F2 status, but no effect of F1 status nor any interaction between F1 status and F2 status (see text). Error bars represent one SEM in each direction.



Discussion

Experiment 3 asked two questions: is F2's perceived responsibility sensitive to the difference between being caused by an intentional versus an unintentional action? No: F2's responsibility ratings did not differ significantly depending on F1's agentive status. These findings rule out a second candidate aspect of agentiveness that might drive interpersonal effect(s) of manipulation: merely intending an immediate action (without intending any of its downstream effects).

Second, is F2's perceived responsibility sensitive to whether F2 acts intentionally versus not? Yes: F2 received higher responsibility ratings when F2 was more agentive.

2.4 Experiment 4: Acting with foresight

Experiment 4 tests the hypothesis that foreseeing the proximal effects of one's action has an effect on responsibility ratings. Experiment 4 builds on Experiment 3, with the more agentive conditions of Experiment 3 (e.g., where the tourist intentionally drops the nut because it's a bad nut, and the businessman intentionally jumps up to avoid being scalded) becoming the less agentive conditions of Experiment 4. The more agentive conditions of Experiment 4 add in that one also foresees the intentional action's proximal effects (e.g., the tourist foresees that the nut will fall into the coffee and cause the businessman to jump up, but only sees these as side-effects of dropping it, and the businessman foresees that jumping up will knock over the table and lead to Cedric's death, but only sees these as side effects of escaping a scalding).

Methods

Participants. Four-hundred-eighty participants were recruited from Amazon Mechanical Turk as in Experiment 1 (mean age = 26, 68.96% female, 121 excluded).

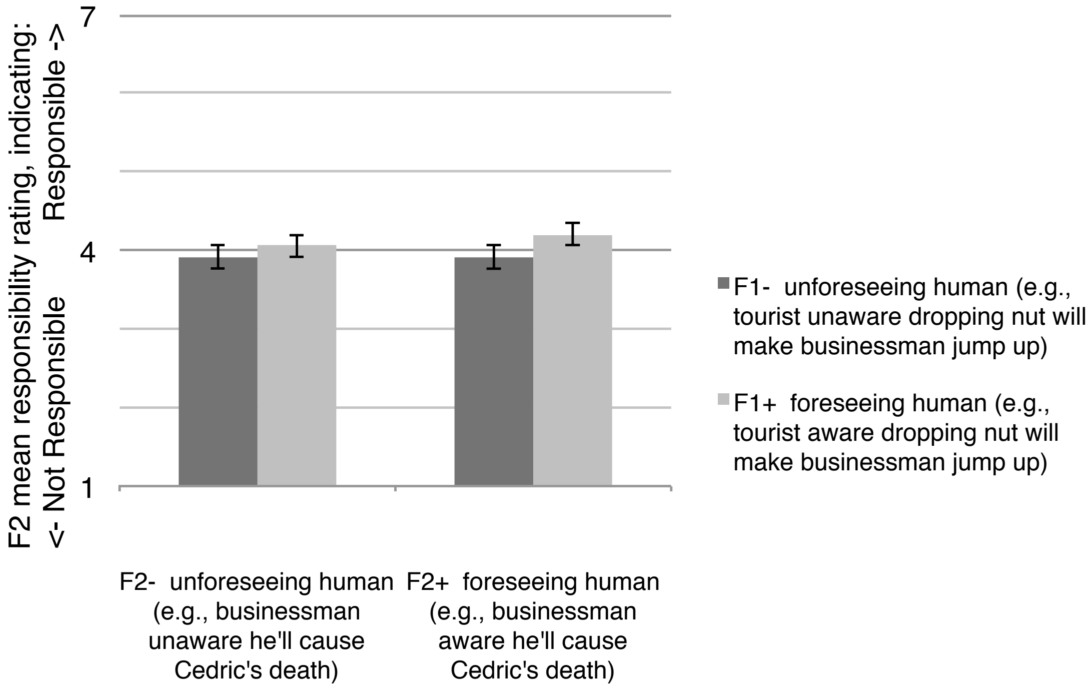
Materials. The four conditions of the six vignettes had the following characteristics: F1 and F2 intend their immediate actions (events 1 and 3, respectively) in all conditions. F1 either does (F1+) or does not (F1-) foresee that F1's action will cause events 2 and 3 (F2's action), and F2 either does (F2+) or does not (F2-) foresee that F2’s action will cause events 4 and 5 (the outcome).

F1 never foresees or intends events 4 or 5 in any condition.

Results

Effects of agency on ratings for F2. A 2x2x6 ANOVA on F2 responsibility ratings with F1 status, F2 status, and vignette as between-subjects factors revealed only a significant main effect of vignette, $F(1, 476) = 34.92$, partial $\eta^2 = .28$, $p < .001$, with no effects of F1 status ($p = .090$), F2 status ($p = .567$), nor an interaction between F1 status and F2 status ($p = .590$).

Figure 5: Effect of F1 status (+, -) on mean F2 responsibility ratings in Experiment 4. There were no significant main effects of F1 status or F2 status, nor any interaction between them (see text). Error bars represent one SEM in each direction.



Discussion

Experiment 4 asked two questions: First, is F2's perceived responsibility sensitive to the difference between being caused by an agent who foresees that her action will affect F2 versus an agent who does not? No: F2's responsibility ratings did not differ significantly depending on F1's agentive status. These findings rule out a third candidate aspect of agentiveness that might drive interpersonal effect(s) of manipulation: foreseeing the proximal effects of one's intended action.

Second, is F2's perceived responsibility for an outcome sensitive to whether F2 foresees it versus not? No: F2's responsibility ratings did not differ significantly depending on F2's agentive status.

2.5 Experiment 5: Intending another's action

Experiments 2-4 ruled out three potential sources of interpersonal effects. Experiments 5-7 now focus on the types of mental states more characteristic of manipulation. Experiments 5-7 each use the same variation in F2 status: F2 either merely foresees that the outcome will be a side effect of F2's action (F2-), as in the more agentive condition of Experiment 4, or F2 also intends the outcome (F2+). For instance, the businessman jumps up intentionally, and either merely foresees that this will knock the table over and lead to Cedric's death as a side effect of avoiding the scalding coffee, or intends thereby to kill Cedric.

However, Experiments 5-7 each vary a different aspect of F1's agentiveness. Experiment 5 begins by testing the specific hypothesis that F1's intention to bring about F2’s immediate action itself, independently of any further effects, affects responsibility attributions to F2. The F1+ condition of Experiment 5 adds (to the more agentive condition of Experiment 4) the intention to bring about the action’s proximal, but no further, effects (e.g., the tourist intends for the nut to fall into the coffee because he wants to fluster the businessman, but without foresight or intent that this will knock the table over or lead to Cedric’s death).

Methods

Participants. Four-hundred-eighty participants were recruited from Amazon Mechanical Turk as in Experiment 1 (mean age = 29, 53.54% female, 180 excluded).

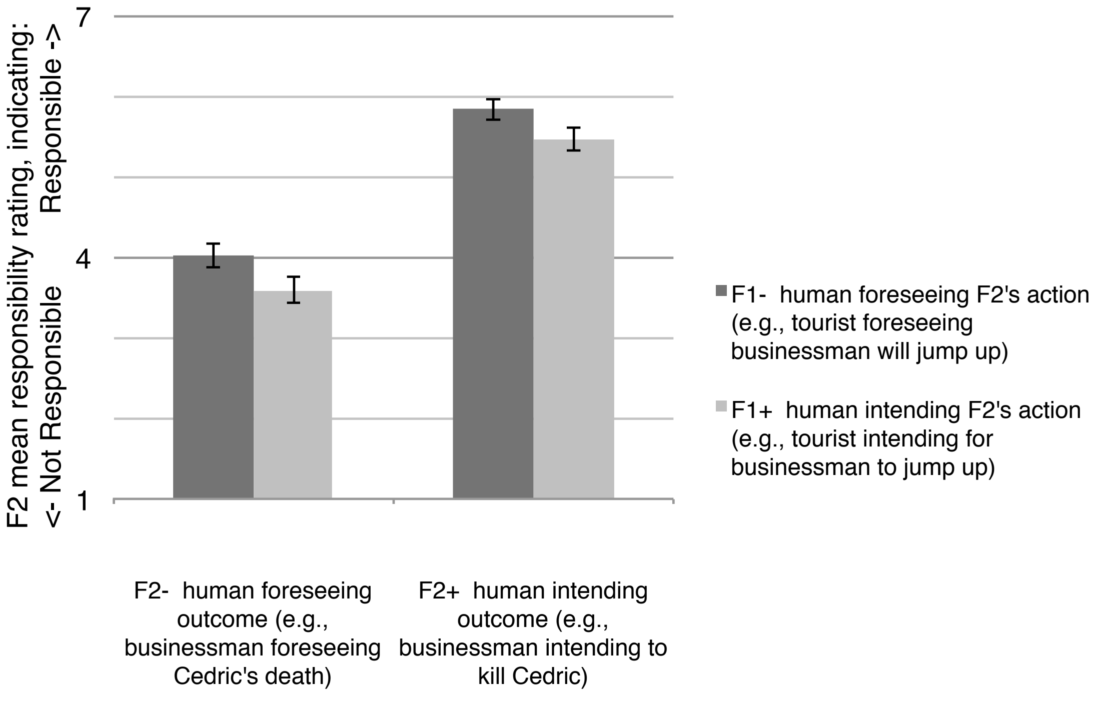
Materials. The four conditions of the six vignettes had the following characteristics: F1 and F2 intend their own immediate actions (events 1 and 3, respectively) and foresee their proximal effects in all conditions (events 1-3 for F1; events 3-5 for F2). F1 additionally intends for F1’s action to cause events 2 and 3 (F1+), or only foresees that it will cause these events (F1-), and F2 either additionally intends for F2's action to bring about events 4 and 5 (F2+) or only foresees that it will (F2-).

Results

Effects of agency on ratings for F2. A 2x2x6 ANOVA on F2 responsibility ratings with F1 status, F2 status, and vignette as between-subjects factors revealed significant main effects of F2 status, $F(1, 476) = 227.39$, $\eta_p^2 = .33$, $p < .001$, with lower ratings when F2 merely foresaw the outcome (F2-) ($M = 3.81$, $SD = 1.74$) than when F2 intended it (F2+) ($M = 5.65$, $SD = 1.52$), and of F1 status, $F(1, 476) = 10.54$, $\eta_p^2 = .02$, p

= .001, with lower ratings for F2 when F1 intended F2's action (F1+) (*M* = 4.53, *SD* = 1.93) than when F1 did not (F1-) (*M* = 4.93, *SD* = 1.80). For example, participants tended to agree that the businessman was less responsible for Cedric's death when the tourist intended the businessman's action than when the tourist merely foresaw it, no matter that the tourist never intended nor foresaw Cedric's death. There was also a significant main effect of vignette, $F(1, 476) = 40.10$, $\eta_p^2 = .31$, $p < .001$, and a significant interaction between vignette and F2 status, $F(1, 476) = 6.00$, $\eta_p^2 = .06$, $p < .001$.

Figure 6: Effect of F1 status (+, -) on mean F2 responsibility ratings in Experiment 5. There were significant main effects of F1 status and F2 status, but no interaction between them (see text). Error bars represent one SEM in each direction.



Discussion

Experiment 5 asked two questions: First, is F2's perceived responsibility sensitive to the difference between a "manipulator" who intends versus merely foresees a particular action of F2's? Yes: F2 received lower responsibility ratings when F1 was more agentive. Experiment 5 is therefore the first in our series to successfully isolate an aspect of agentiveness that drives an interpersonal effect of manipulation: at least in these scenarios, intending to cause the manipulee's action (but none of its effects).

Second, is F2's perceived responsibility for an outcome sensitive to whether F2 intends that outcome versus merely foresees it? Yes: F2 received higher responsibility ratings when F2 was more agentive.

2.6 Experiment 6: Intending the outcome of another's action

Experiment 6 tests the hypothesis that F1's intention to bring about the (same) outcome (as F2) affects responsibility attributions to F2. Experiment 6 builds on Experiment 5. In the F1- condition of the "Manhattan" vignette, for instance, the tourist only intends to fluster the businessman, but foresees the entire chain of events leading to Cedric's death. In the F1+ condition, the tourist also intends to bring about Cedric's death by intervening on the businessman.

Methods

Participants. Four-hundred-eighty participants were recruited from Amazon Mechanical Turk as in Experiment 1 (mean age = 28, 58.13% female, 202 excluded).

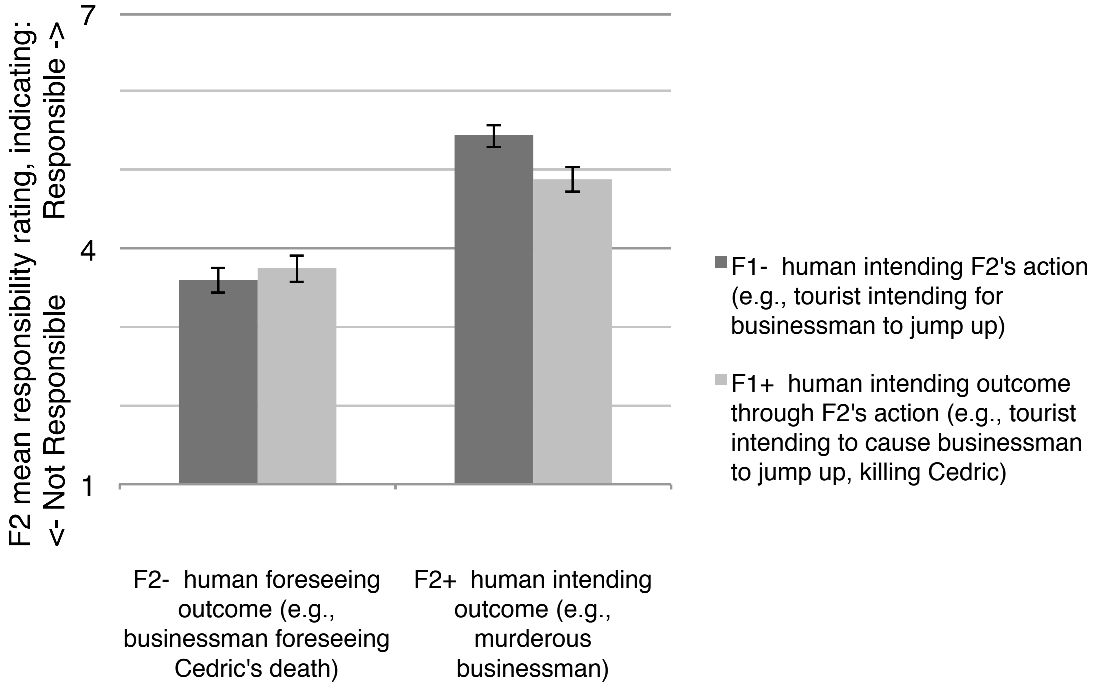
Materials. The four conditions of the six vignettes had the following characteristics: F1 either foresees and intends all events in the causal chain (F1+) or foresees all of these but does not intend events 4 and 5 (the outcome) (F1-), and F2 either foresees and intends events 3-5 (F2+) or foresees all of these events but does not intend events 4 and 5 (F2-). In all conditions, F1 intends to cause F2's action. The only factor varied is whether F1 also intends for F2's action to bring about the outcome (F1+) or not (F1-). Thus, any effect of F1 status on F2 ratings will be due solely to whether or not F1 intends the (same) outcome (as F2), beyond any effect of causing F2's action.

Results

Effects of agency on ratings for F2. A 2x2x6 ANOVA on F2 responsibility ratings with F1 status, F2 status, and vignette as between-subjects factors revealed a significant main effect of F2 status, $F(1, 476) = 122.18$, $\eta_p^2 = .21$, $p < .001$, with higher ratings when F2 intended the outcome (F2+) ($M = 5.15$, $SD = 1.65$) than when F2 merely foresaw it (F2-) ($M = 3.67$, $SD = 1.80$). There was no main effect of F1 status ($p = .126$), but there was a significant interaction between F1 status and F2 status, $F(1, 476) = 7.03$, $\eta_p^2 = .02$, $p = .008$: F1's agentive status had a significant effect on F2 responsibility ratings in the F2+ condition, $t(1, 238) = -2.68$, $p = .008$, with lower ratings for F2 when F1 intended the outcome (F1+) ($M = 4.87$, $SD = 1.74$) than when F1 did not (F1-) ($M = 5.44$, $SD = 1.51$). There was no significant effect of F1 status on F2 ratings in F2- ($p = .521$). For example, the tourist's intending to cause Cedric's death (rather than merely flustering the businessman) had a mitigating effect on the businessman's responsibility, but only when the businessman also intended Cedric's death.

There was also a significant main effect of vignette, $F(1, 476) = 32.79$, $\eta_p^2 = .26$, $p < .001$, and a significant interaction between vignette and F2 status, $F(1, 476) = 3.92$, $\eta_p^2 = .04$, $p = .002$.

Figure 7: Effect of F1 status (+, -) on mean F2 responsibility ratings in Experiment 6. There was a significant main effect of F2 status; while there was no significant main effect of F1 status, there was a significant interaction between F1 status and F2 status, and F1 had a significant effect on F2 responsibility ratings in the F2+ condition (see text). Error bars represent one SEM in each direction.



Discussion

Experiment 6 asked two questions: First, is F2's perceived responsibility for an outcome sensitive to the difference between being manipulated only to act versus being manipulated to bring that specific outcome about? Yes: At least when F2 intended the outcome, ratings for F2 were lower when F1 also intended the outcome than when F1 did not. Experiment 6 is therefore the second to isolate an aspect of agentiveness that drives an interpersonal effect of manipulation.

Second, is F2's perceived responsibility for an outcome sensitive to whether F2 intends that outcome versus merely foresees it? Yes: F2 received higher responsibility ratings when F2 was more agentive.

2.7 Experiment 7: Bypassing

Experiment 7 tests the hypothesis that intentionally *bypassing* F2's normal deliberation and desire-forming processes lowers F2 responsibility attributions. Experiment 7 builds on Experiment 6, with the F1+ condition of Experiment 6 (e.g., where the tourist intends to kill Cedric by dropping a nut, which he foresees will lead the businessman to kill Cedric rather than be scalded by coffee) becoming the F1- condition of Experiment 7. The F1+ condition of Experiment 7 adds in the intention to bring about the outcome through bypassing (e.g., the tourist intends to kill Cedric by dropping a pill containing a mind-controlling drug into the coffee, which he knows will make the businessman want to kill Cedric). This F1+ condition brings us full circle: it's the same F1+ condition used in Experiment 1. The F2 conditions are the same as those in Experiment 5.

Methods

Participants. Four-hundred-eighty participants were recruited from Amazon Mechanical Turk as in Experiment 1 (mean age = 31, 52.08% female, 184 excluded).

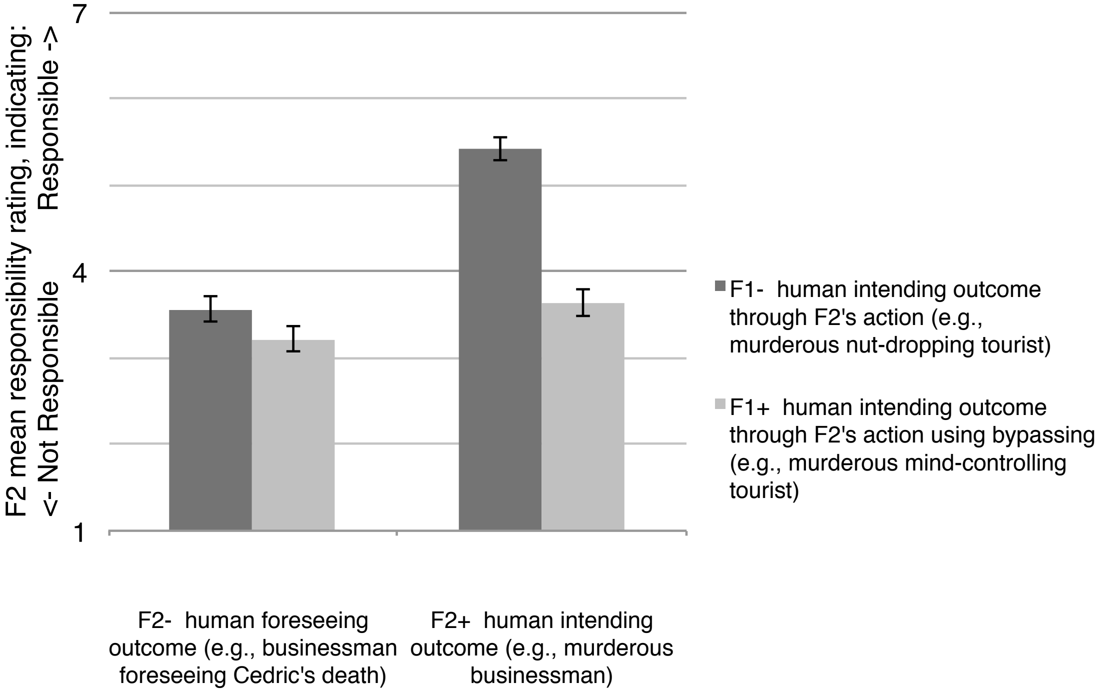
Materials. The four conditions of the six vignettes had the following characteristics: F1 foresees and intends all events in the causal chain and either causes events 3-5 by intervening on F2's external environment (F1-) in a way that F1 knows will lead F2 to reason and act in certain ways, as in Experiments 5 and 6, or by *bypassing* F2's normal reasoning, deliberation, and desire-forming processes through subliminal audio signal, mind-controlling drugs, or similar means (F1+). F2 either foresees and intends events 3-5 (F2+) or foresees all of these but does not intend events 4 and 5 (F2-). Experiment 7 holds fixed that F1 intends to bring about the outcome by causing F2 to intend to cause it, only varying whether F1 does so by bypassing F2's mental states (F1+) or not (F1-). Thus, Experiment 7 tests whether bypassing has an independent interpersonal effect on responsibility attributions, over and above any effects of intending the (same) outcome (as F2) and intending F2's action.

Results

Effects of agency on ratings for F2. A 2x2x6 ANOVA on F2 responsibility ratings with F1 status, F2 status, and vignette as between-subjects factors revealed significant main effects of F2 status, $F(1, 476) = 74.64$, $\eta_p^2 = .14$, $p < .001$, with higher ratings in F2+ ($M = 4.53$, $SD = 1.77$) than F2- ($M = 3.39$, $SD = 1.59$), and of F1 status, $F(1, 476) = 65.61$, $\eta_p^2 = .13$, $p < .001$, with lower ratings given to F2 when F1 caused F2's action through bypassing (F1+) ($M = 3.43$, $SD = 1.60$) than when F1 did not (F1-) ($M = 4.49$, $SD = 1.78$). For example, participants tended to agree that the businessman was less responsible for Cedric's death when the tourist dropped the pill containing the mind-controlling drug, intentionally causing the businessman to kill Cedric by bypassing his mental states, compared to when the tourist merely dropped the nut and intentionally brought about the businessman's action by altering his external environment. There was also a significant interaction between F1 status and F2 status, $F(1, 476) = 29.91$, $\eta_p^2 = .06$, $p < .001$: F1's agentive status had a significant effect on F2 responsibility ratings in the F2+ condition, $t(1, 238) = -9.01$, $p < .001$, with lower ratings for F2 when F1 used bypassing (F1+) ($M = 3.63$, $SD = 1.63$) than when F1 did not (F1-) ($M = 5.42$, $SD = 1.43$). There was no significant effect of F1 status on F2 responsibility ratings in F2- ($p = .091$). For example, the tourist's intention to *bypass* (as opposed to externally manipulate) the businessman's normal deliberation and desire-forming processes had a greater mitigating effect on the businessman's responsibility when the businessman intended Cedric's death than when he merely foresaw it.

Finally, there was a significant main effect of vignette, $F(1, 476) = 13.16$, $\eta_p^2 = .13$, $p < .001$, and significant interactions between vignette and F1 status, $F(1, 476) = 3.07$, $\eta_p^2 = .03$, $p = .010$, and vignette and F2 status, $F(1, 476) = 2.31$, $\eta_p^2 = .03$, $p = .043$.

Figure 8: Effect of F1 status (+, -) on mean F2 responsibility ratings in Experiment 7. There were significant main effects of F1 status and F2 status, as well as a significant interaction between F1 status and F2 status (see text). Error bars represent one SEM in each direction.



Discussion

Experiment 7 asked two questions: First, is F2's perceived responsibility for an outcome sensitive to the difference between being manipulated to bring it about through *bypassing* versus external means? Yes: At least when F2 intended the outcome, ratings for F2 were lower when F1 intended the outcome through bypassing F2's mental states than when F1 only intended the outcome through altering F2's environment. Experiment 7 therefore identifies yet a third aspect of agentiveness that drives an interpersonal effect of manipulation: bypassing.

Second, is F2's perceived responsibility for an outcome sensitive to whether F2 intends that outcome versus merely foresees it? Yes: F2 received higher responsibility ratings when F2 was more agentive.

3. General Discussion

We first discuss the general interpretation of our results and their (philosophical) implications for questions about manipulation before considering additional findings from and limitations of our studies.

3.1 Interpersonal effects on responsibility

We began by considering a philosophical conundrum: does manipulation undermine free will and moral responsibility? If so, is this effect unique to manipulation, or does it extend to cases of deterministic causation? In our initial experiment, we found that canonical cases of manipulation are indeed taken to mitigate responsibility. On the whole, a "manipulee" was judged less responsible when manipulated to perform an action than when caused to perform it by a 'purely deterministic', non-human factor.

Our subsequent experiments revealed which features of canonical manipulation drive this effect. We first ruled out several candidate aspects of agentiveness: being a human being (Experiment 2), merely acting intentionally (Experiment 3), and foreseeing the proximal effects of one's action (Experiment 4). Second, we identified three aspects of a manipulator that *did* mitigate the manipulee's responsibility: intentionally intervening on the manipulee (Exp. 5), intending the final outcome (Exp. 6), and intentionally intervening on the manipulee through bypassing.[8]

Our findings are thus consistent with those of Feltz (2013), Phillips & Shaw (2015), Sripada (2012), and Woolfolk, Doris, and Darley (2006), who all report effects of

[8] When varying whether F1 intends the outcome (Exp. 6), there was no significant effect of F1 status on F2 ratings in the ANOVA, but there was a significant interaction between F1 status and F2 status, and a *t*-test revealed that F1 status had a significant effect on F2 responsibility ratings in the F2+ condition.

manipulation in situations that mirror those from Experiments 5-7, and who suggest that these effects are driven by manipulators' intentions. However, our experimental design also allows us to take two steps beyond this prior research. First, as already noted, our results isolate *which* aspects of manipulators' mental states and intentional influence undermine attributions of responsibility to manipulees: we identify three factors that don't matter (Experiments 2-4) and three factors that do (Experiments 5-7). Second, our experiments address a challenge to extant work that's so far gone unrecognized: previous effects attributed to *manipulation* may stem from a more generic, "symmetric" interpersonal effect, and not from manipulation per se. The effects of manipulators on manipulees observed in prior studies could arise not from any unique causal influence involved in manipulation, but merely because manipulators and manipulees are parts of the same causal chain with partially overlapping actions and intentions.

To motivate this challenge, consider the effect(s) a *manipulee* might have on a *manipulator*'s responsibility. In fact, Woolfolk et al. (2006) not only find that hijackers' intentions affect hostage responsibility ratings, but also that aspects of the *hostage's* agentiveness influence attributions of responsibility to *hijackers*. Even when under the influence of mind-controlling drugs, the extent to which the hostage 'identifies' with his action (roughly, the extent to which it accords with his 'deep self') has an independent effect on the hijackers' responsibility ratings (see also Paharia et al., 2009). Because these effects proceed temporally upstream, from manipulee to manipulator, they cannot be the consequence of any type of (direct) *causal* influence, and so cannot drive any threat of manipulation per se. The existence of such effects raises the possibility that the factors we (and others) have associated with manipulation are not unique to it at all, but constitute a more generic, temporally "symmetric" type of interpersonal effect.

Our data allow us to address this challenge. Recall that in addition to making judgments about the manipulee (F2), we also had participants rate the responsibility of the manipulator (F1), allowing us to investigate which aspects of F2's agentiveness have interpersonal influences on F1. With 2 (F1 status) x 2 (F2 status) x 6 (vignette) ANOVAs commensurate with those above – but using *F1* composite responsibility ratings as the dependent variable – we found significant main effects of F2 status on F1 responsibility ratings in Experiments 1, 5, and 6 (*p*s < .001). Thus, in most cases in which F1's agentive status affected F2 ratings, F2 status also had a significant interpersonal effect in the opposite direction: on F1 ratings. These findings suggest that *intending the same outcome* (even though F2 cannot *cause* F1) lowers attributions of responsibility to F1.[9]

[9] The lack of such an effect in Exp. 7 may be due to a ceiling effect – F1 responsibility ratings were generally quite high; see Table 3.

There were significant main effects of F2 status on F1 responsibility ratings in Experiment 1, $F(1, 476) = 23.85$, $\eta_p^2 = .05$, $p < .001$, Experiment 5, $F(1, 476) = 18.14$, $\eta_p^2 = .04$, $p < .001$, and Experiment 6, $F(1, 476) = 15.53$, $\eta_p^2 = .03$, $p < .001$.

There were significant main effects of F1 status on F1 responsibility ratings in Experiment 1, $F(1, 476) = 1177.25$, $\eta_p^2 = .72$, $p < .001$, Experiment 3, $F(1, 476) = 43.12$, $\eta_p^2 = .09$, $p < .001$, Experiment 4, $F(1, 476) = 11.95$, $\eta_p^2 = .03$, $p = .001$, Experiment 5, $F(1, 476) = 67.68$, $\eta_p^2 = .13$, $p < .001$, Experiment 6, $F(1, 476) = 56.65$, $\eta_p^2 = .11$, $p < .001$, and Experiment 7, $F(1, 476) = 34.08$, $\eta_p^2 = .07$, $p < .001$.

There were also significant interactions between F1 status and F2 status in Experiment 1, $F(1, 476) = 5.40$, $\eta_p^2 = .01$, $p = .021$: the status of F2 had a greater impact on F1 ratings in F1- (a difference of .73 points) than F1+ (a difference of .26 points), and in Experiment 6 $F(1, 476) = 4.11$, $\eta_p^2 = .01$, $p = .043$: F2's agentive status had a greater impact on F1 ratings in F1- (a difference of .61 points) than F1+ (a difference of .20 points). In addition, despite finding neither a significant main effect of F1 status or F2 status in Experiment 2, there was a significant two-way interaction between F1 status and F2 status $F(1, 476) = 6.18$, $\eta_p^2 = .01$, $p = .013$: when F1 was more agentive, F1 was judged more responsible in F2- ($M = 3.55$, $SD = 1.53$) than in F2+ ($M = 3.45$, $SD = 1.49$), but when F1 was less agentive, F1 was judged more responsible in F2+ ($M = 3.70$, $SD = 1.30$) than in F2- ($M = 3.18$, $SD = 1.48$).

Finally, there were main effects of vignette in every study other than Experiments 1 and 7. F1 ratings tended to be higher in "Cliff" than other vignettes. There were also significant interactions between vignette and F1 status in Experiments 2, 3, 5, and 7, between vignette and F2 status in Experiment 1, and between vignette, F1 status, and F2 status in Experiment 3. These were almost uniformly quantitative (rather than qualitative), though unlike with F2 ratings, there were a handful of exceptions: "Amazon" and "Manhattan" in Experiment 2 and "Stable" in Experiment 3 showed a reversed effect of F1 status on F1

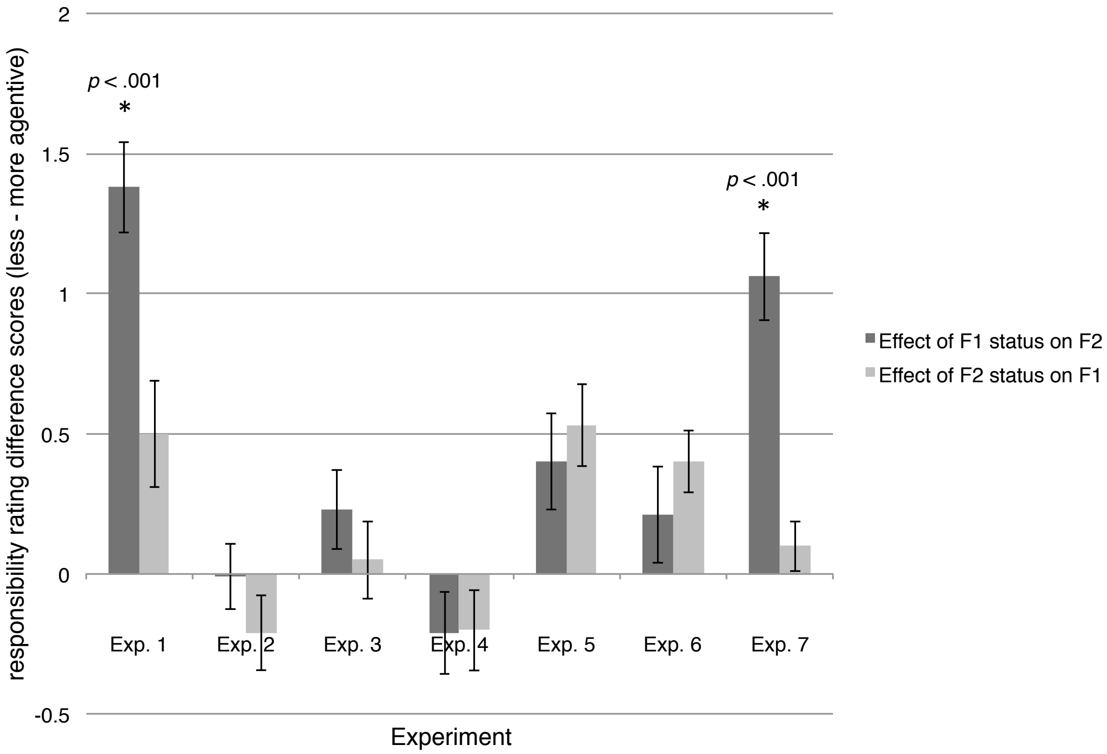
What do these findings reveal about manipulation's *unique* effect(s)? Are the interpersonal effects of manipulators on manipulees completely *symmetrical*, blind to the direction of causation? If so, manipulation arguments might never get off the ground.

Figure 9 compares the relative magnitude of the effect of F1 status on F2 ratings and the effect of F2 status on F1 ratings, presenting the mean *difference* in each factor's responsibility rating as a function of the agentive status of the other factor (e.g., mean ratings for F2 in the F1+ condition subtracted from those for F2 in the F1- condition). The effects of F1 status on F2 ratings were not significantly different from those that F2 had on F1 in Experiments 2-6, but there were large differences in Experiments 1 and 7.[10] Thus, in all and only those experiments that involve *bypassing*, F1's status has a significantly larger effect on F2 than F2 has on F1. We therefore have good evidence that while many interpersonal effects are symmetrical, one uniquely threatening element of canonical manipulation, which must be *asymmetrical*, is bypassing.

Figure 9: Effects of F1 status on F2 responsibility ratings and of F2 status on F1 responsibility ratings in all experiments. Bars represent mean differences between conditions (e.g., F2 responsibility ratings in F1+ subtracted from those in F1-), and thereby represent the magnitude of the effects of F1 status on F2 ratings and vice versa. Cases where the effect of F1 status on F2 ratings was significantly greater than the effect of F2 status on F1 ratings (contrast analysis; see footnote 10) represented by asterisks; error bars represent one SEM in each direction.

responsibility ratings compared to other vignettes, "Manhattan" in Experiment 7 showed no effect of F1 status, and "Cliff" in Experiment 1 showed a reversed effect of F2 status.

[10] To test whether effects of F1 status on F2 responsibility ratings were significantly different from the effects of F2 status on F1 ratings, we conducted a contrast analysis for each study using a univariate ANOVA with 4 conditions: (1) F1+, F2+, (2) F1+, F2-, (3) F1-, F2+, (4) F1-, F2-, with respective contrast weights of +1, -1, +1, -1, and where the dependent measure was computed as F1 ratings minus F2 ratings for conditions 1 and 4, and F1 ratings plus F2 ratings for conditions 2 and 3. Significance in such a contrast analysis would indicate an asymmetry in the effect of F1 status on F2 ratings versus the effect of F2 on F1. This analysis was significant in Experiment 1, $t(1, 476) = 6.42$, $p < .001$, and Experiment 7, $t(1, 476) = 5.71$, $p < .001$, but not in any other experiment (ps > .225).



3.2 Implications for manipulation arguments in philosophy

Our findings put new pressure on philosophical manipulation arguments, which turn on two key premises:

(P1) Manipulation by another agent (such as an evil neuroscientist) undermines an agent's moral responsibility and free will.

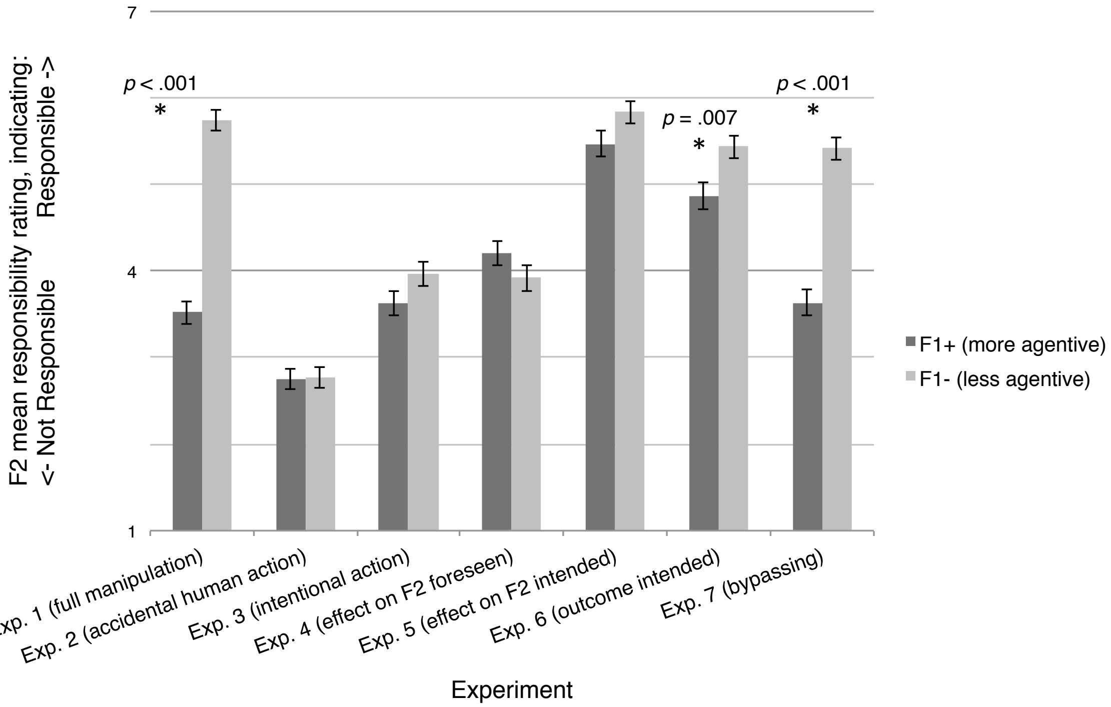
(P2) There is *no relevant difference* between manipulation and matched cases of deterministic causation (e.g., by one's prior mental conditions).

Conclusion: determinism undermines moral responsibility and free will, as well; thus, compatibilism is false.

Sripada (2012) has argued that ordinary intuitions are relevant to manipulation arguments insofar as they're crucial for assessing (P1), which must be specified in terms of some concrete scenario (since, as we've seen, subtly different types of "manipulation" might deliver different verdicts). Whether the manipulee in any given scenario is responsible or not may be intuitively obvious, but precisely which aspect(s) of such scenarios (e.g., which mental states of the manipulator) these intuitions respond to, or track, is not, and is a question for empirical investigation (Sripada & Konrath, 2011).

Consider (P1) in light of our data. Figure 10 presents attributions of responsibility to F2 (when F2 is more agentive) as a function of F1's agentive status. Restricting attention to the subset of cases most relevant to manipulation arguments (those in which F2 intends the outcome – Exps. 1, 5, 6, and 7), the only condition in which (P1) holds true – that is, in which F1's agentiveness leads people to judge that F2 is *not* responsible (as indicated by an average responsibility rating below the mid-point) – is in Experiments 1 and 7: when F1 bypasses F2's mental states. However, if the manipulation in (P1) includes bypassing, then (P2) is false: there *is* a principled difference between that case and the best-matched case of causal determinism. It doesn't appear that both premises of manipulation arguments can be true, at least according to folk intuitions regarding (P1).

Figure 10: Effect of F1 status (+, -) on mean F2 responsibility ratings when F2 is more agentive (F2+). Conditions correspond to those typically considered in manipulation arguments. Cases where F2 ratings differ significantly as a function of F1 status (*t*-tests) represented by asterisks; error bars represent one SEM in each direction.



Deterministic causation, as such, does not involve bypassing. But our results suggest that it's precisely this feature of manipulation that intuitively threatens moral responsibility and free will. Since determinism does not, in principle, share this feature, the comparison to manipulation does not support incompatibilism – deterministic causation and *manipulation* (of the sort that intuitively threatens moral responsibility and free will, at least) *do* appear to be relevantly different.

3.3 Intrapersonal effects of agency on responsibility

Our primary aim was to isolate the factors that drive intuitions about responsibility in cases of manipulation – interpersonal effects – but our findings also shed light on more familiar *intra*personal effects: effects of one agent's mental states on attributions of responsibility to that agent. In particular, we find intrapersonal effects of whether an agent intends her action (Exp. 3), foresees the proximal effects of her action (Exp. 4, though only for F1; see also Lagnado & Channon, 2008), and whether she intends these effects (Exps. 5-7; see also Cushman, 2008 and Young et al., 2008). In all but Experiment 2, F1's ratings were significantly higher in the F1+ than the F1- condition (see footnote 9), and in all but Experiments 2 and 4, F2's ratings were significantly higher in the F2+ than the F2- condition. Thus, the only aspect of agentiveness in our scenarios that did *not* yield an intrapersonal increase in responsibility ratings was whether the factor was a human being as opposed to an inanimate object (Exp. 2).

3.4 Explaining the effects of agency

Having summarized the implications of our findings on attributions of responsibility – including intrapersonal effects, generic interpersonal effects, and the unique interpersonal effect of manipulation – we can consider *why* we observe the particular patterns of attribution we do. Generally, our results suggest that the more events in a causal chain leading to an outcome that one agent intends, the more responsible that agent is judged *and* the less responsible other agents in the causal chain are judged for that outcome. We propose that a unified explanation of these results lies in the *counterfactual robustness* between candidate cause(s) and the effect in question (Hitchcock, 2001; Woodward, 2006), where one agent's intentions can increase the robustness of her counterfactual relation to this effect while decreasing that of other agents involved in the same causal chain.

The idea that attributions of responsibility reflect actual and counterfactual dependence has recent advocates. Gerstenberg and Lagnado (2014) show that responsibility attributions depend not only on the causal relation between an agent and an outcome in the actual world, but also on that relation in other possible worlds. And in explaining why manipulators' intentions influence attributions of blame to manipulees, Phillips & Shaw (2015) appeal to Lombrozo's (2010) *exportable dependence theory*, according to which people judge the statement 'X caused O' appropriate to the extent that (i) had X not occurred in the actual world, O would not have occurred, and (ii) this conditional is *counterfactually robust* (holds in relevant possible worlds).

Counterfactual robustness is especially sensitive to intentions, since agents tend to realize their plans despite variations in the context and means needed to bring them about. If Romeo intends to reach Juliet but is blocked by a wall, he'll find a way around (James, 1890; cf. Heider, 1958). Not so for iron filings blocked from a magnet, or Paris "pursuing" Juliet accidentally. What Romeo has, but Paris (and the iron filings) lack, is a kind of counterfactual control over the outcome: the relation between the outcome and Romeo's mental states is more robust, and less sensitive to contingencies in the causal chain between them. An extension of the exportable dependence theory from causation to responsibility would hold that this type of counterfactual robustness should affect attributions of responsibility, as well.[11]

We can also expect an incremental, "dose-dependent" effect (Campbell, 2006; Woodward, 2010). Roughly, the more events leading up to an outcome an agent intends (and the more fine-grained the description under which she intends them), the stronger the counterfactual robustness between that agent's mental states and the outcome's

[11] McClure et al. (2007) and Lagnado and Channon (2008) suggest that the effect of intentions is not reducible to differences in probability, but they use causal chains devoid of manipulation. In unfolding chains, sufficiency judgments do partially mediate the effect (Hilton et al., 2010). See Supplementary Material B.

occurrence, as well as how it unfolds (this is the scalar version of the point about Romeo and Paris). Hence, the more events an agent intends, the more she'll be held responsible.

The exportable dependence theory correctly predicts the experimental conditions in which we observe intrapersonal effects of agentiveness. As F1 intends more events in the causal chain leading up to the outcome – and so increases the degree of counterfactual robustness between the outcome and her mental states (in Exps. 1, 3, 5, 6, and 7) – F1's responsibility ratings increase.[12] Similarly, as F2 intends more events in the causal chain (Exps. 1, 3, 5, 6, and 7), F2's ratings increase.

The exportable dependence theory also predicts the pattern of interpersonal effects observed in our experiments. In general, the more counterfactually robust the relation between one factor and an outcome, the less robust the counterfactual relation between other factors in the causal chain and that outcome. Intuitively, the more variance in an outcome that one factor accounts for (across possible worlds), the less variance any other factor accounts for. This interpersonal reduction of counterfactual robustness could occur as the result of being caused to bring about the outcome by another agent, rendering it less dependent on one's own desires and deliberation (Exp. 5), or by having one's (normal, non-implanted) mental states bypassed by another agent (Exps. 1 and 7).[13]

[12] Foresight plausibly makes the counterfactual relation more robust, but not to the same extent as intending an event, which would explain the intrapersonal effect of F1 status in Exp. 4. The mob boss' hiring a hitman *even though* he knows it will anger his wife shows he's more willing to perform actions that anger his wife as a side effect in nearby possible worlds (see Uttich & Lombrozo, 2010).

[13] Counterfactual robustness also likely accounts for the interactions between F1 status and F2 status in Exps. 1 and 7. When a causal factor is an accidentally acting robot, animal, or inanimate object (F2- in Exp. 1), it already bears a weaker counterfactual dependence relation with the outcome (compared to when that factor is an agent that intends the outcome (F2+)) and so has less robustness to be reduced by F1. Similarly, because foresight exhibits an intermediate degree of counterfactual robustness, this predicts an intermediate effect of bypassing on F2 ratings when F2 only foresees the outcome (F2- in Exp. 7).

Moreover, merely intending the same outcome as other agents in the causal chain, independent of any causal influence, can also decrease the counterfactual dependence between those agents' mental states and the outcome. Even if Y can't causally affect *X*, Y can still influence the counterfactual robustness of the relation between X and the outcome if Y can causally affect the *outcome*. For instance, hockey player Y might intentionally deflect teammate X's shot into the net. Intuitively, this may reduce X's responsibility for the goal compared to a case in which the deflection is, so far as Y is concerned, completely unintentional. Similarly, hiring a hitman reduces the control the mob boss who hires him has over the hit: it makes it less counterfactually dependent on the mob boss's desires and intentions (the hitman might disobey, for example; see Paharia et al., 2009). This would explain the symmetric interpersonal effect that, following Woolfolk et al. (2006), we observe in our experiments: not only does F1's intending the same outcome as F2 (holding fixed F1's causal influence on F2) reduce responsibility attributions to F2 (Exp. 6), but F2's intending the same outcome as F1 (where F2 can't causally influence F1) also reduces attributions to F1 (Exps. 1, 5, and 6).

In sum, an extension of the exportable dependence theory to responsibility attributions seems to explain all of the observed interpersonal effects of F1's status on F2 responsibility ratings: in Experiments 1 and 5-7 (as F1 intends more events in the causal chain, decreasing the outcome's counterfactual dependence on F2), and all observed effects of F2's status on attributions of responsibility to F1: in Experiments 1, 5, and 6 (which vary whether F2 intends the outcome).[14]

[14] It's plausible that Y's intending outcome O affects X's counterfactual relation to O even if X doesn't intend O. The hitman's intending O makes the counterfactual relation between the mob boss' action (hiring the hitman) and O less robust even if the mob boss' action is somehow accidental (though not as much as if the mob boss' action is intentional; see footnote 13). The exportable dependence theory would then predict

3.5 Additional effects and limitations

So far, we've focused our discussion on inter- and intrapersonal effects of agency on attributions of responsibility. Across experiments, we also found significant effects of vignette. F2 ratings tended to be higher in the "Hospital," "Factory," and "Stable" vignettes than in "Cliff," "Amazon," and "Manhattan" (see Supplementary Material A). In addition, there were significant interactions between vignette and F2 status in several experiments (and between vignette and F1 status in Experiment 7). However, in each experiment where F2 status (or F1 status) had a corresponding significant main effect in the ANOVA, these interactions were only quantitative, not qualitative – that is, F2 status had an effect in the same direction in each vignette, differing only in magnitude.[15] These interactions constitute a potential drawback of our studies, as does the difficulty in matching scenarios along key dimensions without introducing possible confounds or significant narrative implausibility into the vignettes. Despite these limitations, our data suggest that our primary findings prevail across a diverse range of cases.

In addition, our data bear on research in legal scholarship (Hart and Honoré, 1985) investigating interactions between agentiveness and position in causal chains (e.g., being the temporally first versus second cause). We analyze our data in these terms in Supplementary Material B. Our analyses also collapsed across several conceptually distinct judgments: free will, moral responsibility, blame, and causation. This decision was supported by finding high internal consistency between these judgments in all experiments (see footnote 6), but it remains an open question how they might come apart,

the effect of F2 status on F1 ratings in Exp. 5. The failure to find any effect of F2 status on F1 ratings in Exp. 7 may be due to a ceiling effect.

[15] For effects of vignette on F1 ratings, see footnote 9.

whether some mediate others, and whether these judgments are uniformly affected by mental states and counterfactual robustness.

Our studies are also restricted to explicit judgments of responsibility in response to vignettes about third parties. Investigating more implicit and behavioral measures from a broader range of cases is an important direction for future research (though on the general validity of vignette-based measures, see Hainmueller et al., 2015).

Finally, one might worry that the types of bypassing used in our studies – e.g., subliminal audio signal and mind-control – are exceptionally rare. The likelihood of actual bypassing doesn't matter for assessing manipulation arguments and characterizing the cognitive processes involved in responsibility attributions, though. Moreover, even if "direct" bypassing (like hypnosis) is unrealistic, more indirect forms of the kind exemplified by some nudges (Sunstein & Thaler, 2008) may not be. Some even argue that our own unconscious, implicit attitudes may covertly bypass our normal reasoning and desire-forming processes with alarming frequency (Doris, 2015) – precisely the attitudes that nudges are meant to target, and to take advantage of (or "harness").

3.6 Conclusion

Our results corroborate previous findings that one agent's intentions can affect attributions of responsibility to that agent (for a review, see Waldmann, Nagel, and Weigmann, 2012) and to others (Hilton et al., 2005, 2010; McClure et al., 2007; Lagnado & Channon, 2008; Phillips & Shaw, 2015), but they go beyond prior work by isolating which intentions drive this influence. Specifically, we show that three different intentions (investigated in Exps. 5-7) drive three independent interpersonal effects, with the unique threat of manipulation stemming from bypassing.

Our results isolate several types of interpersonal effect, and so provide guidance about what to expect not only in cases of manipulation, but in other types of causal chain that share subsets of its features. Which public policies could threaten free will and moral responsibility? Which technologies exacerbate such threats, and why might divine "manipulation" qualify as an exception? Many of these questions concern interpersonal effects of intentions – and some, perhaps, even bypassing. Our results suggest that, in general, questions surrounding how one agent's responsibility for an outcome are affected by other agents' mental states can be fruitfully guided by asking how they affect the counterfactual robustness between that agent's mental states and the outcome.

Acknowledgments

We're very grateful for the feedback of Neil Levy, Eddy Nahmias, Antonia Peacocke, Derk Pereboom, Jonathan Phillips, Dillon Plunkett, and the referees for *Cognitive Science*. Vanessa Ing assisted with data analysis and collection. This work was partially supported by a James S. McDonnell Foundation Scholar Award in Understanding Human Cognition to T. Lombrozo.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin,* 126, 556–574.

Berlin, I. (1969). Two concepts of liberty. In *Four Essays on Liberty*. Oxford: Oxford University Press.

Carlon, A. (2007). Entrapment, punishment, and the sadistic state. *Virginia Law Review,* 93, 1081–1134.

Cushman, F.A. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition,* 108, 353–380.

Doris, J. (2015). *Talking To Our Selves*. Oxford: Oxford University Press.

Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. (2010). Are your participants gaming the system?: Screening Mechanical Turk workers. *Proceedings of the 28th International Conference on Human Factors in Computing Systems,* 2399–2402.

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

Frankfurt, H. (1969). Alternative possibilities and moral responsibility. *Journal of Philosoph*y, 66, 829–839.

Freedman, D. H. (2012). The perfected self. *The Atlantic,* June 2012.

Gerstenberg, T., and Lagnado, D. A. (2014). Attributing responsibility: Actual and counterfactual worlds. In T. Lombrozo, J. Knobe, and S. Nichols (eds.), *Oxford Studies in Experimental Philosophy,* Vol. 1. Oxford: Oxford University Press.

Guglielmo, S., & Malle, B. F. (2010). Enough skill to kill: Intentionality judgments and the moral valence of action. *Cognition,* 117, 139–150.

Hainmueller, J., Hangartner, D., and Yamamoto, T. 2015. Validating vignette and conjoint survey experiments against real-world behavior. *PNAS,* 112, 2395–2400.

Hart, H.L.A., and Honoré, T. (1985). *Causation in the Law*, 2nd ed. Oxford: Clarendon Press.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. John Wiley & Sons.

Hilton, D. J., McClure, J., and Slugoski, B. (2005). Counterfactuals, conditionals and causality: A social psychological perspective. In D. R. Mandel, D. J. Hilton, & P. Catellani (eds.), *The Psychology of Counterfactual Thinking* (pp. 44–60). London: Routledge.

Hilton, D. J., Mclure, J., and Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology,* 40, 383–400.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy,* 98, 273–299.

James, W. (1890). *The Principles of Psychology.* Cambridge, MA: Harvard University Press.

Johnson, J. T., Ogawa, K. H., Delforge, A., and Early, D. (1989). The effect of position in a causal chain on judgments of legal responsibility. *Personality and Social Psychology Bulletin,* 15, 161–174.

Kane, R. (1996). *The Significance of Free Will*. Oxford: Oxford University Press.

Lagnado, D. A. & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition,* 108, 754–770.

Lagnado, D. A., Gerstenberg, T., and Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science,* 37, 1036–1073.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology,* 61, 303–332.

McClure, J., Hilton, D. J., and Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology,* 37, 879–901.

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

Monroe, A. E. and Malle, B. F. (2009). From uncaused will to conscious choice: The need to study, not speculate about people’s folk concept of free will. *Review of Philosophy and Psychology,* 1, 211–224.

Murray, D., and Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research,* 88, 434–467.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2005). Is incompatibilism intuitive? *Philosophy and Phenomenological Research,* 73, 28–53.

Nahmias, E. & Murray, D. (2010). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff, and K. Frankish (eds.), *New Waves in Philosophy of Action* (pp. 189–216). Palgrave-Macmillan.

Nahmias, E., Shepard, J., and Reuter, S. (2014). It’s OK if ‘my brain made me do it’: People’s intuitions about free will and neuroscientific prediction. *Cognition,* 133, 502–516.

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science,* 331, 1401–1403.

Nichols, S. and Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs,* 41, 663–685.

Paharia, N., Kassam, K. S., Greene, J. D., and Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes,* 109, 134–141.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

Pettit, P. (1997). *Republicanism: A Theory of Freedom and Government*. Oxford: Clarendon Press.

Phillips, J. & Shaw, A. (2015). Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning. *Cognitive Science*, 38, 1320–1347.

Pizarro, D. A. & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (eds.), *The Social Psychology of Morality: Exploring the Causes of Good and Evil* (pp. 99–108). Washington, DC: American Psychological Association.

Shaver, K. G. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness.* New York: Springer-Verlag.

Skinner, B. F. (1948). *Walden Two.* Hackett Publishing Company.

Spellman, B. A., & Kincannon, A. (2001). The relation between counterfactual ("but for") and causal reasoning: Experimental findings and implications for jurors' decisions. *Law and Contemporary Problems: Causation in Law and Science,* 64, 241–264.

Sripada, C. (2012). What makes a manipulated agent unfree? *Philosophy and Phenomenological Research,* 85, 563–593.

Sripada, C. & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind & Language*, 26, 353–380.

Sunstein, C. & Thaler, R. (2008). *Nudge*. Yale University Press.

Uttich, K. & Lombrozo, T. (2010). Norms inform mental state ascriptions: a rational explanation for the side-effect effect. *Cognition,* 116, 87–100.

Waldmann, M., Nagel, J., & Weigmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (eds.), *The Oxford Handbook of Thinking and Reasoning*. Oxford: Oxford University Press.

Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review,* 115, 1–50.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy,* 25, 287–318.

Woolfolk, R. L., Doris, J. M., and Darley, J. M. (2006). Identification, situational constraint, and social cognition: studies in the attribution of moral responsibility. *Cognition,* 100, 283–301.

Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *PNAS,* 107, 6753–6758.

Young, L., Cushman, F., Hauser, M., and Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *PNAS,* 104, 8235–8240.

Young, L. & Phillips, J. (2011). The paradox of moral focus. *Cognition,* 119, 166–178.

Young, L. & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass,* 7, 585–604.