The Folk Psychology of Free Will: Fits and Starts

SHAUN NICHOLS

Abstract: According to agent-causal accounts of free will, agents have the capacity to cause actions, and for a given action, an agent *could have done otherwise*. This paper uses existing results and presents experimental evidence to argue that young children deploy a notion of agent-causation. If young children do have such a notion, however, it remains quite unclear how they acquire it. Several possible acquisition stories are canvassed, including the possibility that the notion of agent-causation develops from a prior notion of obligation. Finally, the paper sets out how this work might illuminate the philosophical problem of free will.

> The language of all mankind, and their ordinary conduct in life, demonstrate, that they have a conviction of some active power in themselves to produce certain motions in their own and in other bodies, and to regulate and direct their own thoughts. This conviction we have so early in life, that we have no remembrance when, or in what way we acquired it (Reid, 1969 [1788], p. 269).

The problem of free will has its roots in commonsense. No scientific expertise is required to bring the problem to life for the uninitiated. For the notion of free will is a commonsense notion, a part of our folk psychology. Over the last quarter century, philosophical work on the notion of free will has flourished (see Kane, 2002 for state-of-the-art overviews). Much of this work is devoted to giving an analysis of the notion of free will that fits with the intuitions of philosophical grown-ups. While this analytic literature on free will has become wildly sophisticated, naturalistic philosophers and developmental psychologists have produced a highly developed literature on the child's emerging capacity for predicting and explaining people's behavior, 'mindreading' (see, e.g. Goldman, 1989; Gopnik and Wellman, 1994; Gordon, 1986; Harris, 1992; Nichols and Stich, 2003; Perner, 1991; Wellman, 1990). Unfortunately, although the notion of free will is indisputably part of folk psychology, the notion of free will has been almost entirely neglected in this naturalistic literature on folk psychology in children. This is especially disappointing for those of us who regard the developmental work as a

I would like to thank the teachers, parents, and children of the N.E. Miles Early Childhood Development Center at the College of Charleston. Thanks to Hilary Martin, Sara Martell, and Emily Askey for assistance in data collection. *Many* thanks to Trisha Folds-Bennett for help getting subjects and pilot data for experiment 2. I would also like to thank John Doris, Larry Krasnoff, Heidi Maibom, Eddy Nahmias, Philip Robbins, Eric Schwitzgebel, two anonymous referees and an editor at *Mind & Language* for helpful comments on an earlier draft.

Address for correspondence: Department of Philosophy, College of Charleston, Charleston, SC 29424, USA.
Email: nichols@cofc.edu

Mind & Language, Vol. 19 No. 5 November 2004, pp. 473–502.

promising source for philosophical illumination. In this paper, I attempt to forge some links between these literatures. Since the issue has not been systematically joined previously, the paper will be rather exploratory and programmatic. I will argue that young children have a notion of agent-causation, according to which (i) actions are caused by agents and (ii) for a given action, an agent *could have done otherwise*. In the first section, I'll set out the notion of agent-causation as it has been developed in philosophy. Then I'll briefly discuss the absence of this notion in standard accounts of the child's understanding of decision making. In section 3, I'll discuss some evidence and present some experimental results that support the claim that children exploit a notion of agent-causation. In section 4, I'll set out the difficult question of acquisition. If children do have a notion of agent-causation, how do they acquire such a thing? I'll describe and discuss several possible acquisition stories. Finally, in the fifth section, I'll consider possible implications for the philosophical debate on free will.

1. The Philosophical Characterization of Agent-Causation

In the philosophical literature on free will, agent-causal theories have a long, if somewhat irregular, history. One of the earliest accounts that is unequivocally agent-causal comes from Thomas Reid (1969 [1788]). According to William Rowe, the foremost contemporary exegete of Reid's view, on Reid's account you count as a true agent-cause of a change in the world when 'you had the power to bring about that change, you exerted that power by acting, and finally, you had the power not to bring about that change' (1989, p. 159). Such agent-causal theories are typically framed against the thesis of determinism, the claim that every event is an inevitable consequence of the prior conditions and the natural laws. Historically, agent-causal theorists have had to fight on two very different fronts. On the one hand, agent-causal theorists are in combat with 'compatibilists' who claim that free will and determinism are perfectly consistent once one gets clear about the proper interpretation of 'free will'. The proper understanding of 'free will', according to agent-causal theorists, is one that is thoroughly at odds with determinism. On the second front, agent-causal theorists resist 'hard determinists', who maintain that while free will is indeed at odds with determinism, that means that free will doesn't exist, since determinism is true. Agent-causal theorists maintain rather that it is a fact that we have free will and that this entails that determinism is false, at least when it comes to agents.[1]

For most of the Twentieth century, agent-causation had little cachet in analytic philosophy. In the mid-Twentieth century, the view was bravely resurrected by

[1] As a result, agent-causal theories count as 'libertarian' approaches to free will. However, some libertarian accounts reject agent-causation (e.g. Ginet, 1990; Kane, 1996).

Roderick Chisholm (1966) and a few others, in the face of rather a good deal of abuse. But recently agent-causation has undergone a renaissance. There are a variety of agent-causal theories, and there has been a resurgence of work on these kinds of accounts (see, e.g. Clarke, 1993, 1996, 2003; Ekstrom, 2000; O'Connor, 1995, 2000; Rowe, 1989, 1991). Despite the important differences between different accounts of agent-causation, all agent-causal theorists agree on two points:

1. An agent is a causal factor in the production of an action.
2. For a given action of an agent, the agent could have *not* caused it. Roughly, the agent *could have done otherwise*.

The first condition has it that the power to make the change, to produce the action, lies in the agent. The important feature here is that agent-causal theories quantify over agents, and they locate in the agent causal powers. However, this first condition alone will not rescue agents from the hegemony of determinism. The problem is that just because the agent caused something to happen doesn't mean that the outcome wasn't inevitable. It doesn't mean that the agent could have refrained from the action. Rowe makes the point by adverting to an example of causal powers outside the realm of agency:

> Suppose a piece of zinc is dropped into some acid, and the acid dissolves the zinc. In this example, we might say that the acid has the power to bring about a certain change in the zinc. . . . But can we reasonably say that the acid had the power not to bring about this change? Clearly we cannot. The acid has no power to refrain from dissolving the zinc . . . The acid, therefore, is not an agent-cause of the zinc's dissolving (Rowe, 1989, p. 159).

The acid has the power to dissolve the zinc, but it does not have the power *not to* dissolve the zinc. One might similarly claim then that an agent was a causal factor in producing an action without maintaining that the agent could have done otherwise than produce the action. Hence, to capture the relevant anti-determinist notion of agent causation, we need to add the condition that the agent *could have done otherwise*, even while all the other factors were exactly the same.[2] The interposed hyphen produces the technical term for this notion, agent-causation.

Agent-causal theories attract skeptics of all sorts. Perhaps the deepest reservation is that agent-causation is unintelligible. For an agent-caused action is produced in a way that is not deterministic, but neither can the action be produced in a *random* fashion. And that, the worry goes, leaves us fresh out of possibilities. I will prescind

[2] In this paper, I intend for the 'could have done otherwise' condition to be understood as indeterminist, unless otherwise noted.

from such questions about whether agent-causation has any remote plausibility. I should probably confess that I am deeply skeptical that we are agent-causes. But here I won't assume or argue that agent-causal theories are wrong. I want to remain entirely neutral on the actual metaphysics of agency. Until section 5, my focus will be on two questions:

i. Do children believe in agent-causation?
ii. How might the belief in agent-causation be acquired?

These questions would not be settled by determining whether or not agent-causation obtains. Children (and adults) believe in all sorts of crazy things, and they fail to believe in all sorts of true things. So we could hardly settle whether children believe in agent-causation by settling the actual metaphysics. Similarly, the belief in agent-causation would demand an acquisition story regardless of whether agent-causation is real or not. For even if it's real, we need a story about how children come to recognize this. Even if we are agent-causes, that does not explain why young children *believe* that we're agent-causes.

2. Mindreading and Agent-Causation

As noted in the introduction, there has been a great deal of work on the child's capacity for mindreading. The research shows that by the age of four, children are remarkably adept at predicting and explaining behavior (e.g. Gopnik and Meltzoff, 1997). This presumably requires a facility at explaining and predicting decisions. Researchers on mindreading have provided increasingly detailed models of children's folk psychological reasoning. So one might expect that this literature is the best place to look to determine whether children believe in agent-causation.

As it happens, on standard accounts of mindreading, agent-causal notions are never invoked. Indeed, models of folk psychology in the mindreading literature make no reference to free will as an element of the lay theory of decision-making. Gopnik and Wellman, for instance, characterize the young child as deploying a version of the practical syllogism:

> 'If an agent desires x, and sees that x exists, he will do things to get x.' Even that form of the practical syllogism is a powerful inferential folk psychological law. It allows children to infer, for example, that if John wants a cookie and sees one in the cookie jar, he will go there for it. If he doesn't want it, or doesn't see it, he won't (Gopnik and Wellman, 1994, p. 265; see also Wellman and Bartsch, 1988; Wellman, 1990).

In our recent work, Steve Stich and I dispute this picture of early mindreading, but we do not invoke agent-causal notions either. Rather, to explain how children predict an agent's decisions, we advert to a 'Mindreading Coordinator' mechanism. This mechanism works as follows on our theory:

> When a prediction about the future behavior of a target is required, the Coordinator's first job is to assemble information about the target's goals or desires.... After this is done, the Coordinator collects whatever information it can find in the Belief Box about the target's desires and sends a call to the Planner [the mechanism for constructing plans]. The Planner's charge is to come up with the best plan for satisfying those desires for an agent in the target's situation. When the Planner reports back with a plan of action, the Coordinator then generates a belief that the target will try to act in accordance with that plan. And it is that belief that is used to predict the target's behavior (Nichols and Stich, 2003, p. 81).

So, on our proposal, the Mindreading Coordinator begins with information about the agent's goals and desires and generates a prediction that the agent will try to act in accordance with the best plan for satisfying his desires. Thus, the Mindreading Coordinator generates a prediction, but it does not generate an attribution of agent-causation.

Similarly, Alan Leslie, who occupies yet another view on the theoretical landscape, does not advert to agent-causal notions as part of the lay view of agency. Leslie maintains that evolution has built us to track three features of agents: Mechanical properties (e.g. 'having an internal and renewable source of energy'), Actional properties ('Agents *act* in pursuit of goals and re-*act* to the environment as a result of perceiving'), and Cognitive properties ('The behavior of Agents is determined by cognitive properties, e.g. holding a certain attitude to the truth of a proposition'). On Leslie's view, these three features of agents are tracked by three different modules that deliver our 'core notions of agency' (Leslie, 1995). Leslie never suggests either that agent-causation is a genuine feature of agents or that agent-causation is a part of our core notions of agency.

Hence, mindreading theorists with very different allegiances share widely in the omission of agent-causal notions from accounts of the lay prediction of behavior. How are we to interpret the evident omission of agent-causation from accounts of mindreading? One possibility is that agent-causation is a wrongly neglected aspect of the mindreading system. Perhaps Gopnik, Wellman, Leslie, Stich, and I have all overlooked a central aspect of the mindreading system, of the mechanisms underlying the child's prediction of behavior. Another possibility, however, is that the notion of agent-causation is not implicated in the mindreading system—in the domain of folk prediction and explanation, the notion of agent-causation plays no useful role, so the mindreading accounts are right to neglect it. An even more skeptical possibility looms. Perhaps the notion of agent-causation has no part in the child's (or the adult's) view of the mind. Perhaps agent-causation is an invention of philosophers that fails entirely to connect with the child's view of decision making. At this point, we arrive at a resolutely empirical issue—do children have a notion of agent-causation?

3. Do Children Believe in Agent-Causation?

As we saw in section 1, there are two crucial claims to an agent-causal theory. First, actions are caused by agents. Second, actions are not inevitable—for a given action the agent could have caused a different action. There is reason to think that children embrace both claims, but making the case for this requires some care.[3] I'll start by arguing that children regard agents as having causal powers to produce actions. Then I'll present some experimental work indicating that children accept that agents *could have done otherwise*.

3.1. Agents Are Causes

Recent work on the understanding of agency in infants and toddlers provides the basis for thinking that young children regard agents as having causal powers. The primary evidence here comes from developmental psychologists, of course. But to make the case for causal powers, we will need to recruit some recent work on causal attribution as well. The goal in this section is to argue that very young children accept the following:

> *The Causal Principle:* An agent is a causal factor in the production of an action; briefly, agents have causal powers to produce actions.

First, I'll rely on developmental evidence to argue that very young children accept the following:

> *The Correlation Principle:* If there is an action, there is an agent.

Then I'll argue that it's plausible that children who accept the Correlation Principle will also accept the Causal Principle.

3.1.1. Developmental Evidence. The available evidence from developmental psychology indicates that infants embrace the Correlation Principle. The first important claim to establish is that infants quantify over agents. As with everything in this paper, the existing data are more suggestive than definitive, but recent work on agency detection provides some reason to think that infants, at least in the second year, do quantify over agents. To begin to see this, we can look to the work of Susan Johnson and her colleagues. In an elegant experiment, 12-month-old infants were shown a fuzzy brown object under a variety of different conditions

[3] Most accounts of children's mindreading probably do maintain that children think that actions are caused by agents. It is the second claim of agent-causal theories that isn't reflected in contemporary accounts of mindreading. However, since the notion of agent-causation encompasses both claims, it will be important to consider directly whether children embrace both claims.

(Johnson, Slaughter and Carey, 1998). In one condition, the fuzzy brown object had eyes, in another condition, the fuzzy brown object was eyeless but interacted contingently with the infant (by beeping and flashing lights when the infant babbled or moved), and in another condition, the fuzzy brown object had neither eyes nor did it interact contingently. In all conditions, children's looking behavior was measured when the fuzzy brown object 'gazed' at one of two objects by making a smooth, 45 degree turn towards the object and remaining in this orientation for several seconds. What Johnson and colleagues found was that infants would follow the 'gaze' of the fuzzy brown object when it had eyes or when it interacted contingently, but infants were less likely to follow the 'gaze' of the fuzzy brown object if it had neither of these features. Johnson proposes that what happened in the experiment is that the infants followed the gaze when the fuzzy brown object was coded as an *agent*, an entity that has mental states (Johnson, 2000, p. 22).[4] For gaze-following is often taken to reflect the 'implicit attribution of a mind to the gazer' (Johnson *et al.*, 1998, p. 233), and the experiments of Johnson and colleagues controlled to exclude many other deflationary interpretations of gaze-following in their tasks (1998, p. 237). Further, they note that experiments on adults using the same fuzzy brown object yielded complementary results. Adults described the fuzzy brown object's turning behavior in mentalistic terms in the same conditions that generated gaze-following in infants (1998, p. 237). So either having a face or interacting contingently apparently triggers a representation of an object as an agent. Further support for the idea that infants code certain objects as agents comes from evidence of diverse behavioral responses to agents. In follow-up work, Johnson, Booth and O'Hearn (2001) explored the reactions of 15-month-old infants to a stuffed orangutan doll that had eyes and exhibited contingent interaction. They found that the infants made communicative gestures to the orangutan, but not to a grossly similar object that lacked eyes and contingent interaction (Johnson *et al.*, 2001, p. 652). Hence, there is some evidence that infants do represent certain objects as agents, insofar as they systematically treat agents distinctively in gaze-following and communication. Furthermore, the work of Johnson and colleagues suggests that the triggers for the representation of agency include having a face and exhibiting contingent interaction. Apparently, then, infants do quantify over agents.

Recall the Correlation Principle, which states that if there's an action, there's an agent. One way to explore whether children accept this principle is by investigating whether children think that if there is no agent, there is no action. Using an ingenious imitation methodology, Meltzoff found that infants attributed goals to people but not to a mechanical device. In Meltzoff's experiment, 18-month-old

[4] Agent-causal theorists often have a stronger notion of agent in mind. Reid, for instance, regarded agents as immaterial substances. The evidence on agency detection doesn't reveal whether or not children share this view of the ontology of agency. And the discussion to follow focuses on a weaker notion of agency according to which agents are entities that have mental states.

infants observed an adult 'failing' to carry out a goal. For instance, in one of the tasks infants observed an adult trying but failing to pull apart a small dumbbell. Infants pulled apart the dumbbell when they watched a person fail, but not when they watched a mechanical contraption fail. According to Meltzoff, the results from this experiment 'demonstrate that physical movements performed by a machine are not ascribed the same meaning as when performed by a person. Therefore, even a weak reading of the data suggests that infants are thinking in terms of goals that are connected to people and not to things' (1995, p. 848). For our purposes, the results suggest the following: the infants do not interpret the machine's behavior as an action because they do not categorize the machine as an agent.

Meltzoff's study suggests that infants think that if there is no agent, there's no action (see also Woodward, 1998), and this supports the claim that infants embrace the Correlation Principle. Further evidence for this claim might be drawn out of the studies by Johnson and her colleagues. As noted above, Johnson and colleagues (1998) found that the infants would follow the 'gaze' of the eyeless fuzzy brown object, but only if it interacted contingently with the infants. If the (eyeless) fuzzy brown object did not exhibit such contingent interaction, the infants would not follow its 'gaze'. One implication of this work is that non-human objects are not categorically excluded from being agents.[5] More importantly for our purposes, this provides further evidence that children accept the Correlation Principle that if there is an action there is an agent. The infants evidently inferred that the fuzzy brown object was an agent from observing the contingent interaction of the fuzzy brown object. A natural interpretation of this is that the infant regarded the contingent interaction as *action*, from which they inferred that there was an agent. That is, the results suggest that babies are inferring that the contingent behavior counts as action and hence, the fuzzy brown object is an *agent*.

Together, the results from Meltzoff and Johnson and colleagues provide promising initial evidence that young children do indeed accept the Correlation Principle that if there is an action, there is an agent. That does not immediately generate the result that children accept the Causal Principle that agents are causes. But now we might appeal to work on causal attribution to build a bridge.

3.1.2. Causal Attribution. The work on causal attribution that I'll invoke is, not surprisingly, quite deliberately non-Humean. Humean accounts of causal attribution eschew causal powers altogether, so it makes no sense on a Humean account of causal attribution to argue that children attribute causal powers to agents or anything else. There is, by now, an important body of empirical work

[5] Indeed, much of the motivation behind the work of Johnson and colleagues is to undermine the claim that infants will only regard humans as agents (Johnson *et al.*, 2001). In the orangutan-doll studies mentioned above, one task is specifically designed to parallel Meltzoff's imitation study, but with an object that is manifestly non-human. Using Meltzoff's method, Johnson and colleagues found that infants would imitate apparently goal-directed actions of the orangutan-doll.

supporting the claim that lay views of causation are decidedly non-Humean (e.g. Ahn and Bailenson, 1996; Bullock *et al.*, 1982; Cheng, 1997; Michotte, 1963 [1946]). Since I'm explicitly trying to see whether children attribute causal powers to agents, I will draw from this non-Humean tradition.

The notion of causal power used in this literature is summarized as follows by Patricia Cheng: 'causal power . . . is the intuitive notion that one thing causes another by virtue of the power or energy that it exerts over the other' (Cheng, 1997, p. 368). Elsewhere, she writes, 'entities and events may have causal powers with respect to other entities or events' (Cheng, 1999, p. 227). Hence, the question for us is whether we can use the fact that children accept the Correlation Principle to infer that children regard agents as entities that have causal powers with respect to actions.

One possibility, of course, is that the Correlation Principle is the product of a prior belief in a version of the Causal Principle. That is, it might be that in the experiments reviewed above, the babies already believe that only agents have causal powers to produce actions, and it is this belief that underlies their responses in the experiments. Indeed, one obvious possibility is that the Causal Principle is innate, that babies are pre-wired to believe that agents are causes.

Even if babies don't have a prior belief in the Causal Principle, there's reason to think that children would likely infer the Causal Principle from the Correlation Principle. On the prevailing account of causal induction, regular covariation is used as evidence for causal induction. There are exceptions to this rule, of course. Sometimes there are spurious correlations. There is currently a flurry of work that tries to build a more tightly constrained account of how correlational information is used to infer causal powers. For instance, Cheng offers an influential account for how subjects estimate causal powers by consulting contingency information (Cheng, 1997, 1999; see also Glymour, 2001). But on all of the prominent accounts (including Cheng's), it is a presupposition of causal attribution that, *ceteris paribus*, if an effect only occurs in the presence of a candidate cause, then that candidate cause does indeed exert causal powers over the effect. To be sure, the promise of these accounts hasn't been fully worked out, and the current formulations will no doubt be subject to counterexamples. But if anything much like the current accounts are right, the Causal Principle would likely be inferred from the Correlation Principle. The Correlation Principle says that effects of the type *action* never occur in the absence of candidate causes of the type *agent*. In that case, given the above presupposition of causal attribution, the Causal Principle would follow. As a result, on this approach, since young children accept the Correlation Principle, if they apply causal reasoning to the domain of action (as seems likely), they will likely come to grant causal powers to agents.

Hence, the developmental evidence suggests that very young children accept the Causal Principle. For the Correlation Principle either reflects a prior belief in the Causal Principle, or, alternatively, the Correlation Principle would likely lead children to infer the Causal Principle. The upshot of this is that the developmental evidence does indeed give us a basis for concluding that young children accept the first condition of agent-causal theories. The developmental research indicates that, from a very young age, children regard agents as distinctive and causal. They quantify over agents, and they locate in the agent a causal power to produce an action.

What we have found thus far is that children apparently attribute causal powers to agents. This alone does not get us to agent-causation. For children might regard agents as causes and also expect such causal activity to fit comfortably into a deterministic framework. Recall the example from Rowe. We might say that the acid has a power to bring about a change in the zinc and that the acid exerted its power. But did the acid have the power *not* to bring about this change? No. That's what excludes it from counting as a true agent-cause. So, what we do not yet have evidence for is the other condition on agent-causal theories: an agent is an agent-cause of a given action only if the agent could have *not* caused it, that is, only if the agent *could have done otherwise*. Neither the evidence on agency detection nor that on causal attribution gives us any reason to think that children accept this condition.

3.2. Agents Could Have Done Otherwise

Although there is excellent work on agency detection, there is, as far as I can tell, no evidence whatsoever on whether children think that an agent could have done otherwise.[6] So I collected data in some very simple experiments. The clearest way to pose the 'could have done otherwise' question is, of course, as a counterfactual. As a result, a rudimentary facility with counterfactuals is necessary for understanding basic questions about whether an agent could have done otherwise. Fortunately, there is good evidence that children do understand counterfactuals from a young age. Three-year olds are good at answering simple counterfactual questions, and four-year-olds can answer even somewhat complicated counterfactual questions (Harris *et al.*, 1996; German and Nichols, 2003). All of the children in the experiments to follow were at least 3-years-old, so it's likely that the children are competent with counterfactuals.

[6] More broadly, there is little evidence on children's understanding of voluntary action. However, there has been intriguing work by Josef Perner on the child's understanding of reflexes and voluntary action. Perner maintains that children don't understand voluntary actions until they have a representational theory of mind (1991), and Perner and Birgit Lang have new results on children's understanding of reflexes and voluntary action. In a group of children aged 3 to 5 years, knee-reflex responses were elicited and the child was told 'Look your leg moved! Did you mean to do this?' Surprisingly, a majority of the children mistakenly said that they did mean to move their leg. Moreover, there was a clear correlation between failing a false belief task and mistakenly judging the knee-reflex to be intentional (Lang and Perner, 2002). The interpretation of this as evidence in favor of Perner's (1991) proposal is complicated by the fact that in some scenarios, children do seem to succeed at recognizing that certain of their actions were not intentional (see Perner, 1991, p. 219). Nonetheless, this is an important area for further research.

Experiment 1

This experiment investigated whether children regard agents as having the capacity to do otherwise. The experiment included two different tasks. In one task, the 'could-have-done-otherwise' task, children were given scenarios in which either a person exhibits some motor behavior (condition 1) or an object moves (condition 2). The child was asked whether the object/person had to behave as it did, or whether it could have done something else instead. In the other task, the 'external constraint' task, children were given scenarios in which a certain possible action was made salient (e.g., taking money off the table), but the action was actually impossible (because the money was glued to the table). In one condition, a character actually tries to perform the action but fails; in the other condition, the character makes no attempt at the action. In each condition, the child was asked whether the character *chose* to act as they did (e.g. leave the money on the table).

Method

Participants

Eighteen children participated. All participants were recruited from the N. E. Miles Early Childhood Development Center at the College of Charleston. The mean age of participants was 4 years, 10.5 months; the range was 3;5 to 6;7. 8 participants were female, 10 were male.

Tasks and materials

Four could-have-done-otherwise items were developed for this experiment. Two of these items probed for whether an *agent* could have done otherwise; the other two items were closely parallel but asked for whether a *thing* could have done otherwise. For instance, in one of the agent cases, children were shown a closed box with a sliding lid. The experimenter said, 'See, the lid is closed and nothing can get in. I'm going to open the lid.' At this point, the experimenter slid the lid open and touched the bottom of the box. Then the child was asked, 'After the lid was open, did I have to touch the bottom, or could I have done something else instead?' In the parallel thing-case, children were shown the closed box with a ball resting on the lid. The experimenter said, 'See, the lid is closed and nothing can get in. I'm going to open the lid.' At this point, the experimenter slid the lid open and the ball fell to the bottom. Then the child was asked, 'After the lid was open, did the ball have to touch the bottom, or could it have done something else instead?'

In addition to these questions, four questions exploring the understanding of choice in the face of *external constraint* were developed. In these questions, modeled on a famous case from Locke (1959 [1689] Book II, chapter 21, section 10), children were presented with scenarios in which an agent is unaware of an external impediment to a certain course of action. In two of these items, the agent does not even attempt to carry out the unavailable action; the other two items were closely parallel except that the agent does attempt to carry out the unavailable action. For instance, in one of the no-attempt questions, children were shown a doll and a small table with pennies glued onto it. The experimenter told the child the following:

> This is Mary. She is walking by this table and sees the money on it. She is trying to decide whether to take the money. Look—the money won't come off the table—it's glued on!—but Mary doesn't know it. Mary doesn't try to pick up the money, so she doesn't know that it won't come off the table. She *thinks* she can take the money off the table. She says, 'I guess I'll leave the money on the table'.

Then the child was asked, 'Did Mary choose to leave the money on the table?' In the parallel attempt-version, the experimenter showed the table with the attached pennies and told the child the following:

> This is Susan. She is walking by this table and sees the money on it. She is trying to decide whether to take the money. Look—the money won't come off the table—it's glued on!—but Susan doesn't know it. Susan tries to take the money and sees that it won't come off the table. She says, 'I guess I'll leave the money on the table'.

The experimenter then asked the child, 'Did Susan choose to leave the money on the table?'.

Procedure

Children were tested individually in a familiar room in their daycare by two experimenters. Two could-have-done-otherwise scenarios and two external-constraint stories were presented to each child. Children were randomly assigned to one of two conditions. Those in condition 1 were presented with the agent-versions of the could-have-done-otherwise questions and with the no-attempt-versions of the external constraint questions. Children in condition 2 were presented with the thing-versions of the could-have-done-otherwise questions and with the attempt-versions of the external constraint questions. There was a similar distribution of ages in each condition. The mean age of participants in the first condition was 4 years, 11 months ($SD = .85$); the mean age of participants in second condition was 4 years, 10 months ($SD = .97$). In each condition, the order of the questions was alternated and counterbalanced. Half of the subjects were given an external constraint question first, the other half were given a could-have-done-otherwise question first.

Results

For the could-have-done-otherwise task, each 'could have done something else' answer was given a score of 1 and the scores were summed, so the cumulative score could range from 0 to 2. Similarly for the external constraint task, each 'yes' was given a score of 1 and the scores were summed. Frequencies of scores are shown in Table 1.

Score	0	1	2
Could have done otherwise			
Agency condition	0	0	9
Thing condition	8	1	0
External constraint			
No attempt	0	1	8
Attempt	4	2	3

Table 1 *Frequencies of scores in experiment 1*

On the could-have-done-otherwise task, there was a significant difference between the conditions ($\chi^2(2,\ N=18)=18.00,\ p<.001$, two-tailed). Children were more likely to say that an *agent* could have done otherwise than that a *thing* could have done otherwise. In the external-constraint task, again there was a significant difference between conditions ($\chi^2(2,\ N=18)=6.6,\ p<.05$, two-tailed). Children in the no-attempt condition were more likely than children in the attempt condition to say that the person chose, e.g. to leave the money on the table. There were no significant correlations between age and responses in any of the conditions.

Discussion

The evidence suggests that children do embrace the second condition on agent-causation theories—the children maintain that an agent *could have done otherwise* than he actually did. The results here were quite strong. In the agent-condition, every single subject said that the agent *could have done something else* on both cases. And in the thing-condition, all but one subject said that the thing *had to* do what it did on both cases.

Of course, the long philosophical tradition of compatibilism immediately brings to mind deflationary interpretations of these results. I can't discuss, much less exclude, all possible deflationary interpretations. Nonetheless, there are two prominent explanations I want to consider. First, one might maintain that the child's notion of choice is merely freedom from external constraint (cf. Hume, 1955 [1743]). So, when the child says that the agent *could have done otherwise* the claim is merely that there is no external obstacle. The external constraint cases were included to explore this possibility. The results on that task suggest that the notion of choice in these children is not simply absence of constraint. For these children, who uniformly judged that the agent could-have-done-otherwise, also tended to think that an agent made a choice even if she was in fact externally constrained from doing otherwise. Hence, on the assumption that the two tasks implicate a single notion of *choice*, that notion is not captured by the freedom-from-external-constraint account of *could have done otherwise*.

Thus, one line of compatibilist interpretation seems not to fit the data. Another, much more ambitious compatibilist proposal is that when people claim that an agent *could have done otherwise*, what they really mean is a tacitly conditionalized 'would' statement: the agent *would have done otherwise* if the conditions had been different. This is the kind of compatibilist response that prompted William James to brand compatibilism a 'quagmire of evasion'. In the present context, the compatibilist might say that when the children claimed that the agent *could have done otherwise*, they were only claiming that the experimenter *would have done otherwise* under different conditions. The contrast with the *thing*-cases goes some distance to rebutting this interpretation. For nearly all the children said that the ball *had* to do as it did, but of course, children understand that the ball *wouldn't* have touched the bottom under different conditions, e.g. if the experimenter had turned the box over as he opened the lid. Nonetheless, the tacit-conditional response is wildly flexible. One might, for instance, appeal to the fact that certain past conditions are easier to undo than others (e.g. Kahneman & Tversky, 1982). So, one might claim that such differences in ease-of-undoing explains the differential effects in the *could have done otherwise* experiment. The flexibility (some might say, the *unprincipled* flexibility) of the tacit-conditional approach makes it impossible to exclude all possible candidate conditions in a case by case manner. To exclude all possible different conditions, one must specify explicitly that the background conditions are held constant. This was done in the next experiment.

Experiment 2

According to one compatibilist interpretation of 'could have done otherwise', when people claim that an agent could have done otherwise, they only mean that the agent would have done otherwise under certain conditions. On this interpretation, *could have done otherwise* is, of course, fully compatible with determinism. To formulate a question that is manifestly incompatible with determinism, one needs to ask whether an agent could have done otherwise, even with all the conditions exactly the same. That was done in the present experiment, which probed intuitions about a physical process and about agents' decisions.

Participants

Nine children participated. All participants were recruited from the N. E. Miles Early Childhood Development Center at the College of Charleston. The mean age of participants was 5 years, 3 months; the range was 4;9 to 5;9. 4 participants were female, 5 were male.

Task and materials

Nine questions were constructed for this task, 3 questions each for the domains of spontaneous choice, moral choice, and physical event. The items also included questions designed to check comprehension. One of the spontaneous choice cases was as follows:

> Scenario: Joan is in an ice cream store and wants some ice cream. She chooses to have vanilla.
>
> Comprehension questions: What happened before Joan chose to have vanilla?

> i. Was Joan in the store before she chose?
> ii. Did Joan decide not to have any ice cream?
> iii. Did Joan want ice cream?
>
> Test question: Okay, now imagine that all of that was exactly the same and that what Joan wanted was exactly the same. If everything in the world was the same right up until she chose vanilla, did Joan have to choose vanilla?

One of the moral choice cases was as follows:

> Scenario: Mary is at a grocery store and wants a candy bar. She chooses to steal the candy bar.
> Comprehension questions: What happened before Mary chose to steal?
>
> i. Was Mary in the store before she chose to steal?
> ii. Did Mary leave the store before she stole the candy bar?
> iii. Did Mary want a candy bar?
>
> Test question: Okay, now imagine that all of that was exactly the same and that what Mary wanted was exactly the same. If everything in the world was the same right up until she chose to steal, did Mary have to choose to steal?

Finally, one of the physical event cases was as follows

> Scenario: A pot of water is put on a stove and heated up. The water boils.
> Comprehension questions: What happened before the water boiled?
>
> i. Was the pot on a stove before the water boiled?
> ii. Was the stove turned down before the water boiled?
> iii. Did the water get hotter?
>
> Test question: Okay, now imagine that all of that was exactly the same. If everything in the world was the same right up until the water boiled, did the water have to boil?

All of the children got most of the comprehension questions right. When a comprehension question was missed, the scenario was repeated followed by the comprehension check again.

Procedure

Children were tested individually in a familiar room in their daycare by two experimenters. All children were given all 9 cases. The presentation of the cases was alternated by domain and counterbalanced for order.

Results

Each 'yes' answer was given a score of 1 and the scores were summed for each domain (spontaneous, moral, physical), so the cumulative score could range from 0 to 3. Mean scores and standard deviations are given in Table 2.

	Score (SD)
Had to happen	
Spontaneous choice	1.00 (0.50)
Moral choice	0.67 (1.38)
Physical event	1.56 (0.88)

Table 2 *Mean scores for experiment 2 (SD in parentheses)*

There was a significant difference in responses between the physical cases and the moral cases ($t(8) = -4.438$, $p < .01$, two-tailed). Participants were more likely to say that the outcome had to happen for the physical cases than for the moral cases. The pattern of responses here is telling. All of the children treated at least one physical case as deterministic. Two children gave deterministic responses to all 9 items. Crucially, though, the rest of the children gave indeterminist responses to all 3 of the moral cases. Interestingly, there was no significant difference between physical and spontaneous cases ($t(8) = -1.474$, $p = .179$, n.s.). This might simply be because the sample is small, but it is clearly an issue for further study.[7]

Given the results of the above experiments, I suggest, as a tentative hypothesis, that children regard agents as having the capacity to have done otherwise in a way that can't merely be reduced to a conditionalized analysis. The above experiments hardly rule out all the possible alternative interpretations. No small set of experiments could do that. But the evidence certainly fits the indeterminist interpretation. And there's not (yet) any evidence against it. On the contrary, the available evidence provides support for the claim that children embrace both claims of the agent-causal account. Apparently children think that an agents is a causal factor in the production of an action. They also seem to think that when an agent produces an action, he could have done otherwise, and moreover, that

[7] A significant difference between spontaneous cases and physical cases was found in two pilot studies with adults. Participants were instructed to indicate (on a 6-point scale) the extent to which they agreed with statements like 'If everything was exactly the same up until the moment the water boiled, then the water had to boil at that moment'. In the first pilot study, participants (N = 16) were more likely to disagree with the claim that the man had to choose vanilla ($M = 2.19$, $SD = 1.22$) than with the claim that the water had to boil ($M = 3.69$, $SD = 2.09$), ($t(15) = 2.90$, $p < .05$, two-tailed). In the second pilot study, participants (N = 34) were again more likely to disagree with the claim that the man had to choose vanilla ($M = 3.50$, $SD = 1.24$) than with the claim that the water had to boil ($M = 4.18$, $SD = 1.38$), ($t(33) = 2.34$, $p < .05$, two-tailed); they were also more likely to disagree with the claim that a person had to choose to steal ($M = 3.35$, $SD = 1.25$) than with the claim that the water had to boil ($t(33) = 2.80$, $p < .01$, two-tailed). The rather different means in these two pilot studies, as well as the high standard deviations, indicates that there is considerable individual variation on these questions for adults.

when an agent makes a moral choice, he was not determined to choose as he did.

4. How Do Children Acquire the Belief in Agent-Causation?

Let's suppose that the above proposal is right, that children believe in agent-causation. If that's right, it leaves us with a vexing question—how do children acquire this notion of agent-causation? What is especially challenging is to determine how children come to believe in the indeterminist could-have-done-otherwise principle. For this principle makes the notion of agent-causation notoriously difficult to fit into any standard metaphysics (see, e.g. Strawson, 1986; van Inwagen, 1998). Here I will consider several possible acquisition stories. My preferred candidate story will come last. But, I should note up front, I know of no clear evidence against any of the proposals. As a result, this is not intended as a complete or remotely conclusive discussion of acquisition. My goal is primarily to get clearer about the some of the options and directions for future work.

4.1. Learning, 3rd Person

One possible acquisition story, flowing out of the empiricist tradition that so richly informs our scientific thought, is that the child infers that we are agent-causes by theory building (cf. Gopnik and Meltzoff, 1997). To use Paul Harris' memorable phrase, this approach views the child as a 'stubborn autodidact' (Harris, 2002). To develop such a learning story, one might begin by recruiting a Cheng-style account of causal inference to argue that children infer that agents are causes (as sketched in section 3.1.). That's the easy part. The difficult residual question is how children acquire the other part of the agent-causal account. That is, how do children infer that an agent *could have done otherwise*? Few developmental theorists, I suspect, would want to maintain that people *are* agent-causes, and that children accordingly build the right theory. And it's not clear what general principles would lead to learning something like this. Here I will only consider one flat-footed possibility. Perhaps children use a general inference rule of the following form:

Phenomena that routinely and persistently defy predictability are indeterministic. That is, children might, when faced with unpredictable phenomena, infer an indeterministic process as the underlying explanation. If children embrace such a principle, they might use the fact that their behavior predictions are very often flouted to conclude that agents produce decisions in a way that is indeterministic. However, this runs up against the fact that lots of processes children witness are thoroughly unpredictable. Indeed, as compatibilists like to point out, agents seem to be no more unpredictable than the weather. But as far as we can tell, children do not infer an indeterministic process as the underlying explanation for the vicissitudes of the weather. Perhaps children do believe in indeterminism about such matters, but one would like to see a more systematic empirical approach to the issue before throwing in with the sort of learning theory sketched above.[8]

4.2. Learning, First Person

While contemporary developmentalists in the learning tradition tend to focus on how children learn from observing the external world, in the case of agent-causation, one might expect that it's more plausible to look *inside* to find the source for inferring that one is an agent-cause. Agent-causal theorists themselves often incline toward a first-person learning story for how we come to believe in agent-causation. For instance, Reid suggests that the notion of agent-causation is acquired from experience of our own choices:

> It is very probable, that the very conception or idea of active power, and efficient causes, is derived from our voluntary exertions in producing effects; and that, if we were not conscious of such exertions, we should have no conception at all of a cause, or of active power (Reid, 1969 [1788], Book IV, chapter 2).[9]

More recent agent-causal theorists also suggest that our experience of our choices provides some kind of evidence for the existence of agent-causation. Indeed, agent-causal theorists typically take this as the primary evidence that we are agent-causes. Campbell writes:

> The appeal is throughout to one's own experience in the actual taking of the moral decision as a *creative* activity in the situation of moral temptation. 'Is it possible', we must ask, 'for anyone so circumstanced to *dis*believe that he could be deciding otherwise?' The answer is surely not in doubt. When we decide to exert moral effort to resist a temptation, we feel quite certain that we *could* withhold the effort; just as, if we decide to withhold the effort and yield to our desires, we feel quite certain that we *could* exert it—otherwise we should not blame ourselves afterwards for having succumbed (Campbell, 1957, p. 169).

Here is O'Connor:

> The agency theory is appealing because it captures the way we experience our own activity. It does not seem to me (at least ordinarily) that I am caused to

[8] There is work on the understanding of 'randomness' in children, but the research really probes for the child's understanding of distribution phenomena in probabilistic events. For instance, in Piaget and Inhelder's famous 'marble tilt box'-task, children are asked to predict the arrangement of colored marbles after the box has been tilted (Piaget and Inhelder, 1975; see Metz, 1998 for discussion of more recent work). But this work does not attempt to plumb whether children attribute to physical events the kind of *deep indeterminism* required for agent-causation.

[9] Reid goes on to suggest that this acquisition story is also the only way to explain our belief in causation in the physical world. But that additional proposal is regarded as excessive by contemporary agent-causal theorists (see O'Connor, 2002, p. 345).

> act by the reasons which favor doing so; it seems to be the case, rather, that *I* produce my decision *in view of* those reasons, and could have, in an unconditional sense, decided differently . . . Just as the non-Humean is apt to maintain that we not only perceive, e.g., the movement of the axe along with the separation of the wood, but the axe *splitting* the wood . . . , so I have the apparent perception of my actively and freely deciding to take Seneca Street to my destination and not Buffalo instead (O'Connor, 1995, p. 196–7).

In the above passages, Campbell and O'Connor suggest that our experience of our decision making provides some evidence for the truth of agent-causation. Of course, we might dispute the reliability of such introspective evidence, but we have set aside issues about whether agent-causation obtains. What we want to know is whether first person experience serves as the basis for acquiring the notion of agent-causation. And even if introspective evidence is deeply suspect, it might nonetheless be the basis for our belief in agent-causation.

When agent-causal theorists appeal to experience, they provide precious few details about how the experience provides evidence for agent-causation. Thus, Campbell says 'we feel certain that we could withhold the effort'; O'Connor writes, 'it seems . . . that I produce my decision . . . and could have . . . decided differently'. The important gap is that they don't explain why the introspective experience provides evidence that we *enjoy* agent-causation as opposed to evidence that we *believe* that we enjoy agent-causation. For our purposes, of course, this is a vital difference. For we are already allowing that we do believe that we enjoy agent-causation. The question at hand is *why* do we believe this. How might experience be a source of evidence (even problematic evidence) that we *enjoy* agent-causation?

It's important to emphasize at this point that even if agent-causation is real, even if our actions actually spring from agent-causation, that alone wouldn't explain why our experience of our choices leads us to believe in agent-causation. We still need a story about how the experience we have would lead us to believe in agent-causation.

One possibility is that experience delivers a raw feel, a *quale*, of agent-causation. On this proposal, we come to believe in agent-causation for the same sort of reason we come to believe in pain. It's experientially immediate and primitive. But this just seems phenomenologically implausible, at least to me (I suppose I can't speak for you). The conviction of agent-causation doesn't seem to come from anything so phenomenologically simple. What we do have is a *belief* that we could do otherwise. And it's not implausible that this belief infuses our experience of our own choices. But in that case, we still need a story about how we acquired the belief in the first place.

There is, I think, a way to move towards an introspection-based learning account of the acquisition of the notion of agent-causation. In recent work, Stephen Stich and I have argued that while introspection does provide access to one's current mental states, introspection does not provide access to the theoretical and practical reasoning processes that underlie one's decisions (Nichols and Stich, 2002). As a result, on this account, while I know what my decisions are by direct introspection, I lack such access to the machinations that produce those decisions. Hence, one might maintain that I feel like I could have decided otherwise because my experience fails to reveal any deterministic underpinnings of my decision making.[10]

While this provides the beginning of an acquisition account, there is a major lacuna. For simply because we lack access to any deterministic decision making process does not, by itself, explain why we would think that our decisions are *not* produced by a deterministic process. After all, we don't perceive the causal mechanisms subserving sun spots, but that doesn't lead us to doubt that there is a causal-deterministic story to be told. Indeed, this is true even for some behaviors. We typically don't have access to the underpinnings of bodily tics, but it's not at all clear that we infer that there is no deterministic story to be told. On the contrary, at least as adults, we suppose that there is some deterministic story to be told about the causal underpinnings of bodily tics. So the claim that we don't perceive a deterministic process of decision making must be supplemented to explain the intuition that our decisions are not determined.

One way to supplement the introspective account is to maintain that people have an abiding belief that we *do* have access to all the causal factors and the causal processes underlying our own decision making. If people do believe in such introspective transparency, then it would be appropriate, given the above facts, for people to infer that one could have done otherwise. For if one introspects no deterministic process underlying one's decision making and one also thinks that if there *were* a deterministic process, one *would* introspect it, one could infer that there is no deterministic process.

To say that we have a belief in introspective transparency is, of course, an empirical claim. In the present context we want to explain the young child's acquisition of the belief in agent-causation, so the relevant empirical claim would have to be that preschool children have an abiding belief in introspective transparency. *Do* preschool children have an abiding belief that they have access to all the mental processes that occur? I realize that this riff is getting tiresome, but we just don't have any evidence on the matter. There certainly is no evidence in favor of the claim that children have such an abiding belief. Furthermore, even if children believe both that they enjoy introspective transparency and that they introspect no deterministic process underlying their decision making, it's a further hypothesis that they draw an inference over these two beliefs to arrive at the notion of agent-causation. Thus we lack appropriate evidence to be enthusiastic about this learning story too.

[10] Philip Robbins has reminded me that something like this proposal was offered by Paul Holbach, an Eighteenth century French hard determinist. Holbach maintained that we mistakenly believe in free will because we fail to perceive the complicated mechanistic factors that actually produce our behavior (Holbach, 1970 [1770], chapter XI). Of course here the agent-causal theorist will insist that there *is* no deterministic process that produces our behavior. But again I'm trying to be neutral on the actual metaphysics. My claim here is meant to preserve the neutrality—whether there's a deterministic process or not, we don't have introspective access to one.

4.3. Nativism

The previous approaches share the assumption that the belief in agent-causation is a product of inference over evidence. At the other pole, one might simply embrace a resolute nativism about the notion of agent-causation. Nativism is no newcomer to discussions of folk psychology. Indeed, some of the most prominent theorists of folk psychology maintain that vital parts of folk psychology are innate (e.g. Fodor, 1992; Leslie, 1994). Hence, it is certainly a live theoretical option to press for a nativism here as well. In this context, then, a nativist might propose that the agent-causal notion of choice is part of our innate folk psychological endowment.

In section 3.1.2, I suggested that the work on agency detection might be taken to indicate that the child has innate knowledge that agents cause actions.[11] Some of the work on agency detection certainly sits well with the core knowledge program (see Carey and Spelke, 1996). According to the core knowledge approach, the infant comes pre-loaded with crucial innate structures, including innate notions of object and number, and these core elements persevere into adulthood. A core knowledge theorist might well embrace the idea that the notion of agent-causation is an element in the core knowledge package. Just as there's reason to think that infants have an innate notion of physical cause (see, e.g. Leslie, 1995), there might be reason to posit another innate notion of causation—agent-causation. Indeed, there is a nice parallel here. For theorists who embrace non-Humean notions of causation outside of the domain of agency often maintain that the non-Humean notion of causal powers must be innate. As a result, there is a satisfying symmetry in maintaining that this other notion of causal power—agent-causation—is also innate.

4.4. From Obligation to Agent-Causation

A somewhat different approach to the issue of acquisition is to consider features of the child's psychology that would *facilitate* the acquisition of the notion of agent-causation (cf. Sperber, 1996). One key element in the cognition of young children seems particularly promising for developing such an acquisition story—from a young age children have a vivid notion of obligation. The basic story that I'm inclined towards accordingly owes a perverse debt to Kant. For I'm attracted to the idea that the belief in agent-causation derives in part from a prior belief that people have obligations. According to standard interpretations, Kant (1956 [1788]) thinks that the existence of free will can be inferred from the fact that we ought to follow the moral law together with the fact that 'ought' implies 'can' (see e.g., Guyer, 1998; Scruton, 1982). Kant maintains that it is a fact that we are obligated to act in accordance with the moral law. From this we can derive, according to Kant, another fact—that we are (indeterministically) free. For we can't be obligated to do the impossible, and if determinism is

[11] In this context, the term 'knowledge' is not supposed to carry an implication of truth. Rather, innate knowledge is a body of information that might well be false.

true, it is impossible for us ever to do other than we are determined to do.[12] This ingenious line of argument might be deployed in a naturalistic environment, divested of any commitments on the actual metaphysics of morals.[13]

The Kantian argument suggests, then, the following acquisition hypothesis:

> Children come to believe in agent-causation as a result of a prior belief in obligation.

To defend this hypothesis would involve showing that:

(i) children apply a notion of obligation, and
(ii) this notion of obligation implies *could have done otherwise* (in an indeterminist sense).

Defending the first claim is easy. From a young age, children do have a notion of obligation. Indeed, from a young age, children are adept with several notions of 'ought'. They recognize conventional 'oughts', e.g. 'he shouldn't have yelled in class', moral oughts, e.g. 'she shouldn't have pulled Mary's hair', and prudential oughts, e.g. 'he shouldn't have climbed that tree'.[14] Children think that people *ought* to adhere to such rules, and in some cases, flouting the rule should be met with punishment. Furthermore, in the case of moral obligations, children think that people ought to adhere to certain forms of behavior even if there is no rule-giver (see e.g. Smetana, 1993). Like Kant, children regard moral obligations as independent of external rule-givers and as generalizable (see Nichols, 2004 for discussion). The experimental work indicates that children have a notion of obligation (or perhaps several different notions) quite early, by the age of 2 or 3 (see Harris, 2000 for review). Some have suggested that there are innate mechanisms devoted to the notion of obligation, and there are a number of different adaptationist proposals on offer for the evolutionary origins of such capacities (see e.g. Cosmides, 1989; Cummins, 1996, 1998; Sripada and Stich forthcoming). Alternatively, the child's early grasp of the notion of obligation might be learned, drawing on the massive exposure children have to normative prohibitions (Harris, 2000, pp. 156–7).

[12] The debt is perverse because Kant does not maintain that the moral law plays a role in the acquisition of the belief that we are free. Rather, the moral law plays a role in the *proof* that we are free. Larry Krasnoff has pointed out to me that there are passages in the *Groundwork* in which Kant might be interpreted as offering his own account of how people come to believe that they are agent-causes (Kant, 1964 [1785], pp. 447–449).

[13] That is, since we are not trying to prove the existence of indeterminist free will, we need not be committed either to the claim that it's a fact that we have moral obligations or to the claim that we genuinely have indeterminist free will. For our goal is only to try to figure out how children acquire the belief that choice is indeterminist. To serve that goal, again, the true metaphysics is not the issue. Rather, what matters is what children believe. In the present context, then, what matters is not whether it is a fact that we have moral obligations but only whether children believe that we have moral obligations.

[14] Obviously, I'm using 'ought' very loosely, but greater precision isn't necessary given our goals here.

Children get an enormous amount of verbal information about obligations. They are told, from a young age, that they *shouldn't* do this, that they *ought to* do that. They are also told, after performing an action, that they *shouldn't* have done it, and punishments are routinely meted out accordingly.

The Kantian argument is that since it's a fact that one ought to follow the moral law, and it's a fact that 'ought' implies free will, it follows that one has free will. At this point, we have reason to think that children do have a notion of obligation. Indeed, children clearly think that a person *ought* to behave morally. The difficult question at this juncture is whether the young child's notion of *ought* does imply *could have done otherwise.*[15] Children certainly say that a child who hits another child *shouldn't* have done it.[16] They say that a naughty child *ought to* have behaved (see, e.g., Harris, 2000, pp. 157ff). In Paul Harris' important treatment of obligation, he explicitly argues that the child's notion of obligation can be reduced to a conjunction of the notion of free agency and a notion of goal-directedness: 'children's understanding of obligation . . . amounts to a conjunction of two different concepts: the concept of a free agent who can either carry out or withhold a given action, and the concept of a goal-directed agent whose action leads to a particular outcome' (Harris, 2000, p. 159). So, on Harris' view, the young child's notion of *ought* does plausibly imply free will, in the sense reflected by the principle that an agent *could have done otherwise.*

Of course, a compatibilist might offer a sophisticated interpretation of 'ought' on which 'ought' does not imply *could have done otherwise*. However, in the present context, we're already granting that children have the metaphysically strange notion of agent-causation. This drains much of the motivation behind adopting compatibilist interpretations. In any case, there's little reason to think that fathers and mothers or, for that matter, Mother Nature would take care to instill a notion of obligation that is judiciously hedged so that it is weak enough not to imply *could have done otherwise*. It's hard to see why the pressures (whether parental or evolutionary) that lead to a notion of obligation would be constrained away from agent-causal implications.

Even if children do have a notion of obligation that carries the implication of *could have done otherwise*, this alone does not guarantee that children would come to believe in agent-causation. For we don't draw out all the implications from our childhood beliefs within our lifetimes, much less before first grade. Nonetheless, the preceding considerations about obligation indicate that the child's psychological space would be heavily primed to adopt this belief. The notion of

[15] Again, it's possible that children have multiple different notions of obligation. But what matters for present purposes is only whether they apply *at least one* notion of obligation that implies *could have done otherwise*. I'll focus on moral obligation, following Kant's lead, but the acquisition story might not require the idea of full strength moral obligations. Conventional or prudential obligations might suffice.

[16] I can't resist recounting an anecdote: in pilot studies for experiment 2, one child said 'He *shouldn't* have stole the candy so he didn't *have* to steal it.'

agent-causation would be strongly attractive in a psychological space in which the notion of obligation figures prominently.

The attractiveness of the idea of agent-causation would only be enhanced by the considerations about introspective access noted in 4.2,. Introspectively, we lack access to the processes that eventuate in our decisions. In particular, we get no introspective evidence of deterministic decision making processes. As a result, we are not assaulted with evidence that flouts the idea that we could have done otherwise. This might further facilitate the fixation of the belief that we are agent-causes.

Before closing this section, I want to return to Harris' remark that the child's concept of a free agent forms an important part of the child's concept of obligation. Harris and I share the suspicion that in children there is a deep connection between the notions of obligation and free will. However, of course, what Harris is really suggesting here is the exact contrary of what I am proposing. Harris wants to explain the acquisition of the concept of obligation, so he appeals to a prior concept of free will. I want to explain the acquisition of the notion of free will (as agent-causation), so I appeal to a prior notion of obligation. We both want to pass the buck, only in opposite directions. One consideration in favoring of passing the buck from freedom to obligation is that we have a few promising proposals concerning the notion of obligation. By contrast, as yet we don't have any detailed stories about the acquisition of the notion of free will in children. Hence the current paper.

As advertised, this roundup of acquisition stories is neither decisive nor complete. Although I find the obligation-based approach appealing, this is not because I have a shred of evidence against other proposals. Rather, I find the obligation-based approach attractive because there is good reason to think that some of the crucial elements are in place. Children clearly have a notion of obligation; this notion is plausibly a notion that implies *could have done otherwise*; and the lack of introspective evidence of deterministic processes would facilitate the belief that an agent could have done otherwise.

5. Philosophical Implications

I'd now like to turn to some even more speculative matters—how the preceding discussion might bear on the philosophical problem of free will. One way to think about the problem of free will is that it is driven by powerful intuitions that push in different directions. On the one hand, we have an intuition of indeterminism with respect to agency. This, of course, has been our focus for the bulk of the paper, and it plays a central role in philosophical arguments for agent-causation (Campbell, 1957; O'Connor, 1995; Reid, 1969 [1788]). However, the opponents of agent-causation often appeal to a different set of intuitions about agency. According to these philosophers, we have a strong intuition of determinism with respect to agency. This point is made with characteristic flair by Hume:

> It is universally acknowledged that there is a great uniformity among the actions of men, in all nations and ages, and that human nature remains still the same, in its principles and operations. The same motives always produce the same actions: The same events follow from the same causes . . .
> it appears, not only that the conjunction between motives and voluntary actions is as regular and uniform as that between the cause and effect in any part of nature; but also that this regular conjunction has been universally acknowledged among mankind, and has never been the subject of dispute, either in philosophy or common life (Hume, 1955/1743, pp. 83, 88).

Hume's point is that we accept determinism, not just with respect to billiard balls and poisons, but also with respect to agents.

If we attempt to give a psychological explanation of these intuitions, a natural proposal is that there are two different systems underlying the opposing intuitions. But such a proposal will seem *ad hoc* without some kind of support. The developmental work might eventually contribute some of this support. For the experiments reported in this paper, together with the previous work in mindreading, suggest the following as a working hypothesis: there are two (at least partly) independent systems of agency attributions. One system, the mindreading system, generates deterministic agency attributions. Here I've suggested that the notion of agent-causation might be delivered by a different system, the obligation system. Of course, this is all entirely tentative, but it will be useful to explore how such a two-system account might be developed.

In section 2, we saw that contemporary accounts of mindreading neglect indeterminist intuitions. But as noted earlier, this neglect might be wholly appropriate. For it's plausible that in the domain of folk prediction and explanation, the notion of indeterminism plays no useful role. Of course folk psychological predictions often go wrong, and this might be acknowledged within the system itself. But there is no need to invoke indeterminism to explain inaccuracies. For we might simply suppose that the principles guiding folk psychological prediction and explanation are, like the principles guiding much folk biological prediction and explanation, hedged with *ceteris paribus* clauses. If so, then the notion of indeterminism makes no contribution to the predictive or explanatory power of the mindreading system. This point is only reinforced if we think of the origins of the mindreading system. The capacity for mindreading is plausibly a product of our biological evolutionary heritage (e.g. Leslie, 1995; Nichols and Stich, 2003). Being able to predict the behavior of other animals accurately would carry great fitness advantages, and there is some reason to think that our mindreading mechanisms have precursors in nonhuman animals (e.g. Hare *et al.,* 2000). Again, in this context, adding a notion of agent-causation to the mindreading system contributes no additional predictive power, so there would be no call for Mother Nature to graft this into the system.

The situation is quite different when we turn to the obligation system, and I want to sketch a very crude account of the potential utility of the notion of agent-causation for obligation systems. The first point to make is that in contrast to the mindreading system, the function of obligation systems is not to predict behavior. Instead, it's plausible that a primary function of obligation systems is to regulate behavior, to generate *different* behavior.[17] The notion of agent-causation might make a significant contribution to the function of behavior regulation. The issues here get sticky quickly, but let's allow that the obligation system is committed to some version of the maxim that 'ought' implies 'can'.[18] Even compatibilists often want to admit *some* version of this maxim (though, of course, compatibilists have distinctive views about how to interpret the 'can'; see e.g., Frankfurt, 1988; Railton, 1995; Slote, 2001). Furthermore, it's common ground that the maxim is used to excuse a person's apparent transgressions. The obligation to choose to do *X* won't apply when agents *can't* choose to do *X*. In this light, it's clear that agent-causal theories severely limit the range of excuses. For on agent-causal theories, when it comes to choice, it will nearly always be the case that we could have done (i.e. chosen) otherwise. So the excuse that 'I couldn't have chosen otherwise' will basically never be available. Why might this enhance the obligation system's function of behavior regulation? Because, the proposal continues, *typically*, the effectiveness of an obligation system at regulating behavior will diminish as the number of viable excuses increases. For as the excuses proliferate, the obligations lose their reach.[19] If, in making our assessments of obligations, we assume that in nearly every case, the person has the (agent-causal) power to choose in accordance with his obligations, then an enormous range of excuses is immediately ruled out. This would allow us to wield our normative cudgels much more widely, and perhaps thereby to mould behavior more effectively.

Obviously, the foregoing is only the most programmatic proposal. A great deal of work would be required to show empirical merit in the above sketch of the origins of our intuitions about agency. However, the philosophical interest of pursuing the project is, I hope, evident. For it seems quite possible that the conflicting intuitions that drive the philosophical problem of free will derive

[17] One obvious question here is what notion of function is being invoked. The proposal might be run with different notions of function. For instance, one might approach this issue in terms of evolutionary biology and maintain that natural selection generated obligations systems to serve the function of behavior regulation. However, one might take a less biological approach and maintain that obligations systems are produced by cultures partly in order to serve the function of behavior regulation.

[18] Clearly this raises a further question—why is the *'ought' implies 'can'* maxim part of the obligation system? As far as I know, this question has never been broached in a naturalistic setting, and I will not be able to take it up here.

[19] There are all sorts of complications that this hypothesis must address. For instance, does the hypothesis apply even when it systematically excludes excuses that are actually appropriate? This issue arises acutely in the present case. For if determinism is true about psychological processes (as many think), then many excuses of the form 'I couldn't choose otherwise' might actually be apt. If the excuses are apt, is it still the case that a system that excludes those excuses will be more effective at regulating behavior? A hard question, to be sure, but it's by no means obvious that the answer is 'no'.

from different cognitive systems. That might explain the historical recalcitrance of the philosophical problem of free will.

6. Conclusion

In this paper I've argued that the notion of agent-causation is plausibly part of the young child's folk psychology. I've also set out a range of proposals for how this notion might be acquired. I've made some efforts to try to establish my favored views, but I harbor no illusions that I've presented a complete case. I do, however, harbor the hope that there will in future be greater consideration of how the notion of free will fits into the child's early understanding of the mind. The philosophical dividends of the developmental work might be substantial indeed. Kant famously maintained that there are unassailable arguments both for and against causal determinism. It remains a possibility, I think, that such a tension is present within folk psychology itself and that folk psychology simply does not present us with a consistent metaphysics of agency. That is, it's possible that folk psychology, construed broadly, is committed both to agent-causation and to a deterministic story about psychological processes. I've tentatively suggested that these opposing commitments derive from two quite different systems that produce agency attributions. Evaluating this proposal is an empirical enterprise that will require a much richer account of the underpinnings of the commonsense understanding of free will.

Department of Philosophy
College of Charleston

References

Ahn, W. and Bailenson, J. 1996: Causal attribution as a search for underlying mechanisms. *Cognitive Psychology*, 31, 82–123.

Bullock, M., Gelman, R., and Baillargeon, R. 1982: The development of causal reasoning. In W. Friedman (ed.), *The Developmental Psychology of Time*. New York: Academic Press, 209–254.

Campbell, C. A. 1957: *On Selfhood and Godhood*. London: George Allen & Unwin.

Carey, S. and E. Spelke. 1996: Science and core knowledge. *Philosophy of Science*, 63, 515–533.

Cheng, P. 1997: From covariation to causation: A causal power theory. *Psychological Review*, 104, 367–405.

Cheng, P. 1999: Causality in mind: Estimating contextual and conjunctive power. In F. Keil and R. Wilson (eds.) *Explanation and Cognition*. Cambridge: MIT Press.

Chisholm, R. 1966: Freedom and action. In Keith Lehrer (ed.), *Freedom and Determinism*. New York: Random House. Pp. 11–44.

Clarke, R. 1993: Toward a credible agent-causal account of free will. *Noûs*, 27, 191–203.

Clarke, R. 1996: Agent Causation and Event Causation in the Production of Free Action. *Philosophical Topics*, 24, No. 2, 19–48.

Clarke, R. 2003: *Libertarian Accounts of Free Will.* New York: Oxford University Press.

Cosmides, L. 1989: The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31, 187–276.

Cummins, D. 1996: Evidence of Deontic Reasoning in 3- and 4- Year Old Children. *Memory and Cognition*, 24, 823–829.

Cummins, D. 1998: Social norms and other minds. In D. Cummins and C. Allen, *The Evolution of Mind.* Oxford: OUP.

Ekstrom, L. 2000. *Free Will: A Philosophical Study*. Boulder: Westview Press.

Fodor, J. 1992: A theory of the child's theory of mind. *Cognition*, 44, 283–296.

Frankfurt, H. 1988: What we are morally responsible for. In his *The Importance of What We Care About.* Cambridge: Cambridge University Press.

German, T. and Nichols, S. 2003: Children's counterfactual inferences about long and short causal chains. *Developmental Science*, 6, 514–523.

Ginet, C. 1990: *On Action*. Cambridge: Cambridge University Press.

Glymour, C. 2001: *The Mind's Arrows*. Cambridge, MA: MIT Press.

Goldman, A. 1989: Interpretation psychologized. *Mind & Language*, 4, 161–185.

Gopnik, A. and Meltzoff A. 1997: *Words, Thoughts and Theories.* Cambridge, MA: MIT Press.

Gopnik, A. and Wellman, H. 1994: The theory theory. In L. Hirschfeld and S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture.* New York: Cambridge University Press. Pp. 257–293.

Gordon, R. 1986: Folk psychology as simulation. *Mind & Language*, 1, 158–170.

Guyer, P. 1998: Kant, Immanuel. In E. Craig (ed.) *Routledge Encyclopedia of Philosophy*, .

Hare, B., Call, J., Agnetta, B, and Tomasello, M. 2000: Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, 59, 771–785.

Harris, P. 1992: From simulation to folk psychology: The case for development. *Mind & Language*, 7, 120–144.

Harris, P. 2000: *The Work of the Imagination.* Oxford: Blackwell Publishers.

Harris, P. 2002: What do children learn from testimony? In P. Carruthers, S. Stich and M. Siegal (eds.), *The Cognitive Basis of Science.* Cambridge: Cambridge University Press. Pp. 316–334.

Harris, P., German, T. and Mills, P. 1996: Children's use of counterfactual thinking in causal reasoning. *Cognition*, 61, 233–259.

Holbach, P. 1970 [1770]: *The System of Nature: Or, Laws of the Moral and Physical World.* Translated by H. D. Robinson. New York, B. Franklin.

Hume, D. 1955 [1743]: *An Enquiry concerning Human Understanding.* L. Selby-Bigge (ed.). Oxford: Clarendon Press.

Johnson, S. 2000: The recognition of mentalistic agents in infants. *Trends in Cognitive Sciences*, 4, 22–28.

Johnson, S. Booth, A. and O'Hearn, K. 2001: Inferring the goals of a nonhuman agent. *Cognitive Development*, 16, 637–656.

Johnson, S., Slaughter, V. and Carey, S. 1998: Whose gaze will infants follow? Features that elicit gaze-following in 12-month-olds. *Developmental Science*, 1, 233–238.

Kahneman, D. and Tversky, A. 1982: The simulation heuristic. In D. Kahneman, P. Slovic, and A. Tversky (eds.) *Judgment Under Uncertainty*. Cambridge: Cambridge University Press.

Kane, R. 1996: *The Significance of Free Will*. New York: Oxford University Press.

Kane, R. (ed.) 2002: *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Kant, I. 1956 [1788]: *The Critique of Practical Reason*. Trans. L. Beck. Indianapolis: Bobbs-Merrill.

Kant, I. 1964 [1785]: *Groundwork of the Metaphysics of Morals*. Trans. H. Paton. New York: Harper & Row.

Lang, B. and Perner, J. 2002: Understanding of intention and false belief and the development of self control. *British Journal of Developmental Psychology*, 20, 67–76.

Leslie, A. 1994: ToMM, ToBY and agency: Core architecture and domain specificity. In L. Hirschfeld and S. Gelman (eds.) *Mapping the Mind*. Cambridge: Cambridge University Press. Pp. 119–148.

Leslie, A. 1995: A theory of agency. In D. Sperber, D. Premack and A. Premack (eds.) *Causal Cognition*. Oxford University Press.

Locke, J. 1959 [1689]: *An Essay concerning Human Understanding*. A. Fraser (ed.). New York: Dover.

Meltzoff, A. 1995: Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31, 838–850.

Metz, K. 1998: Emergent understanding and attribution of randomness. *Cognition and Instruction*, 16, 285–365.

Michotte, A. 1963 [1946]: *The Perception of Causality*. New York: Basic Books.

Nichols, S. 2004: *Sentimental Rules: On the Natural Foundations of Moral Judgment*. New York: Oxford University Press.

Nichols, S. and Stich, S. 2002: How to read your own mind: A cognitive theory of self-consciousness. In *Consciousness: New Philosophical Essays* Q. Smith and A. Jokic (eds.). Oxford University Press, Pp. 157–200.

Nichols, S. and Stich, S. 2003: *Mindreading*. Oxford: Oxford University Press.

O'Connor, T. 1995: Agent causation. In T. O'Connor (ed.), *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. New York: Oxford University Press. Pp. 173–200.

O'Connor, T. 2000: *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.

O'Connor, T. 2002: Libertarian views: Dualist and agent-causal theories. In R. Kane (ed.) *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Perner, J. 1991: *Understanding the Representational Mind*. MIT Press.

Piaget, J. and Inhelder, B. 1975: *The Origin of the Idea of Chance in Children*. London: Routledge & Kegan Paul.

Railton, P. 1995: Made in the shade: Moral compatibilism and the aims of moral theory. In J. Couture and K. Nielsen (eds.) *On the Relevance of Metaethics: New Essays on Metaethics, Canadian Journal of Philosophy*, Supplementary Volume 21, pp. 79–106.

Reid, T. 1969 [1788]: *Essays on the Active Powers of the Human Mind.* Cambridge, Massachusetts: MIT Press.

Rowe, W. 1989: Two concepts of freedom. *The Proceedings and Addresses of the American Philosophical Association*, 61, 43–64. Reprinted in O'Connor (ed.), *Agents, Causes, and Events: Essays on Indeterminism and Free Will.* New York: Oxford University Press, 1995. (All page references are to the reprinted article.)

Rowe, William L. 1991: *Thomas Reid on Freedom and Morality.* Ithaca: Cornell University Press.

Scruton, R. 1982: *Kant.* Oxford: Oxford University Press.

Slote, M. 2001: *Morals from Motives.* New York: Oxford University Press.

Smetana, J. 1993: Understanding of social rules. In M. Bennett (ed.) *The Development of Social Cognition: The Child as Psychologist.* New York: Guilford Press. Pp. 111–141.

Sperber, D. 1996: *Explaining Culture.* Cambridge, Mass: Blackwell.

Strawson, G. 1986: *Freedom and Belief.* Oxford: Clarendon Press.

Van Inwagen, P. 1998: The mystery of metaphysical freedom. In P. van Inwagen and D. Zimmerman (eds.) *Metaphysics: The Big Questions.* Oxford: Blackwell Publishers. Pp. 365–373.

Wellman, H. 1990: *The Child's Theory of Mind.* MIT Press.

Wellman, H. and Bartsch, K. 1988: Young children's reasoning about beliefs. *Cognition*, 30, 239–277.

Woodward, A. 1998: Infants selectively encode the goal object of an actor's reach. *Cognition*, 69, 1–34.