Predictability, Causation, and Free Will*

Luke Misenheimer (University of California Berkeley)

August 18, 2008

The philosophical debate between compatibilists and incompatibilists about free will and determinism is old and complicated, and both sides might seem to have many advantages and disadvantages. As Nahmias, Morris, Nadelhoffer, and Turner point out in their recent paper "Is Incompatibilism Intuitive?" (2006), many incompatibilists claim their side has the advantage of according with the ordinary intuitions of ordinary people. Nahmias et al. set out to investigate this claim systematically by giving ordinary people questionnaires, each with a brief vignette describing an action in a deterministic universe and then a question either about whether the agent in question acted of his or her own free will or about whether the agent in question is fully morally responsible for the action. They carried out several studies, each with a different questionnaire, but all their results seem to indicate one thing: a strong majority of ordinary people are actually *compatibilists* about free will and determinism.

In response, Nichols and Knobe (2007) carried out studies of their own, also using questionnaires that first presented a deterministic universe and then asked about moral responsibility. However, although their methods were in many important ways similar to the methods of Nahmias et al., Nichols and Knobe seemed to get very different results. The results of the Nichols and Knobe studies seem to indicate that a strong majority of ordinary people are incompatibilists after all.

*I wrote this paper as part of my honors thesis at the philosophy department of UNC Chapel Hill. I had a *great* deal of help at all stages of the project from Joshua Knobe, my thesis advisor. I also received helpful comments on the project from John Roberts, Jesse Prinz, Ram Neta, and Shaun Nichols.

The apparent disagreement between the Nahmias et al. studies and the Nichols and Knobe studies leads us to wonder what difference between the questionnaires leads to such a striking difference in results, especially as knowing this difference could help us determine how each groups' studies are relevant – or irrelevant – to the philosophical debate. Each group claims that only its own studies are philosophically relevant, explaining away the other group's results by pointing to some aspect of their questionnaire other than the determinism described or the free will asked about (Nahmias, 2006; Nichols and Knobe, 2007). However, it might be that *both* groups' studies are philosophically relevant and that there is a difference between the groups' questionnaires beyond the superficial differences with which they both attempt to explain away each other's results. I conducted a study of my own to see whether a particularly interesting difference in the questionnaires might make a real difference in ordinary judgments about free will.

1 Background

To see several potentially important differences, we will examine one condition in a study by Nahmias et al. and one condition in a study by Nichols and Knobe. In the condition from Nahmias, et al. (2006), subjects were told to imagine a particular deterministic universe:

> Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict everything about how it will be with 100% accuracy.

Subjects were then told about a particular act that this supercomputer predicts: Jeremy Hall robs a bank. When asked about Jeremy's moral responsibility, 83% of subjects gave the *compatibilist* response.

In the condition from Nichols and Knobe (2007), subjects were told to imagine another deterministic universe:

> Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it had to happen that John would decide to have French Fries.

Subjects were later told about a particular act in Universe A: Mark cheats on his taxes. When asked about Mark's moral responsibility, 77% of subjects gave the *incompatibilist* response.

There are several clear differences between the studies. One is that the Nichols and Knobe vignette includes talk about necessity that the Nahmias et al. vignette does not include. This is the difference that Nahmias (2006) proposes explains the difference in the studies' results. According to the Nichols and Knobe vignette, "if everything in the universe was exactly the same up until John made his decision, then it *had to happen* that John would decide to have French Fries." This could be taken to mean the same as "if everything in the universe was exactly the same up until John made his decision, then *necessarily* John would decide to have French Fries." This reading differs in the scope of the necessity operator from what would be implied by determinism: "(nomologically) *necessarily*, if everything in the universe was exactly the same up until John made his decision, then John would decide to have French Fries." If, thanks to this difference in the scope of the modal operator, subjects in the Nichols and Knobe studies take all actions in Universe A to be metaphysically – or even nomologically – necessary, then their denial of free will is unsurprising and reveals very little about the relationship between their concepts of free will and determinism.

Another difference between the studies is that the Nahmias et al. vignette includes talk about a potentially violent crime that the Nichols and Knobe vignette does not include. This is the difference that Nichols and Knobe (2007) propose explains the difference in the studies' results. In the Nahmias et al. vignette, Jeremy Hall robs a bank. Robbing a bank is a serious crime, sometimes – at least in movies – resulting in some bodily harm or restriction of autonomy to bank employees and customers. This aspect of the crime might cause subjects to become angry at Jeremy Hall upon reading that he commits it. Subjects' anger might then cause them to strongly desire to blame Jeremy Hall. This desire might override subjects' conscious reasoning, which would normally lead them to deny that Jeremy robs the bank with a free will.

Even though both these differences in the vignettes might be contributing a great deal to the differences in subjects' responses, neither points to something really interesting about ordinary people's *concepts of free will and determinism.* There is another difference between the studies, though, that *might* point to something more interesting: the difference between predictability and causation. The determinism in the Nahmias et al. study is formulated in terms of perfect predictability based on pre-birth events, and the determinism in the Nichols and Knobe study is formulated primarily in terms of complete causation by pre-birth events.

If this difference between the formulations of determinism in the two studies makes a difference in subjects' judgments about free will, the two studies have revealed a very interesting facet of the ordinary concepts of free will and determinism, viz. that the ordinary concept of free will is compatible with determinism but *not with causation*, or at least not with causal determinism. It might seem strange that causation would conflict with free will but predictability would not; after all, how can perfect predictability be explained but by inferring complete causation? Just how causation might conflict with the ordinary concept of free will is an interesting topic in itself, but before investigating *how* it conflicts we need to investigate *whether* it conflicts. To discover whether the difference between predictability and causation does indeed make a difference in ordinary judgments about free will, I had to

run a study of my own.

2 Experiment

My study was designed simply to discover whether the difference between complete causation and perfect predictability makes a difference for judgments of free will and what kind of a difference it makes. Subjects in my study were told either about an imaginary universe in which pre-birth events completely cause all an agent's actions or about an imaginary universe in which a description of the birth-time world enables perfect prediction of all an agent's actions. Subjects were then asked whether a particular action in this universe could have been performed with a free will.

2.1 Methods and secondary results

Subjects in my study were 60 willing passersby on the UNC north campus. They were randomly assigned to one of two conditions: the causation condition and the predictability condition. They were given a questionnaire with a brief vignette about an imaginary world populated by imaginary beings called "Nuhams". The vignette began the same way in both conditions:

> Imagine another universe that is like ours in almost every way. In this universe there is a planet that is very much like Earth, and on this planet there are beings very much like us. They call themselves Nuhams. Some Nuhams are teachers, some are construction workers, some are criminals, some are scientists, and many others have many other jobs, just like us.

In the causation condition, subjects learned that every action of any given Nuham is completely caused by events before the Nuham's birth:

> However, Nuham scientists have recently discovered a very interesting fact about their world. Everything that any given Nuham does is *completely caused* by

> things that happened before it, which are themselves completely caused by things that happened before them, which are themselves completely caused by things that happened before them ... and so on back to things which happened before the Nuham was born. So everything a Nuham does is the result of a chain of complete causation going back to before that Nuham was born.

Whereas in the predictability condition, subjects instead learned that every action of any given Nuham is perfectly predictable based on a description of the world at the time of the Nuham's birth:

> However, Nuham scientists have recently discovered a very interesting fact about their world. Based on a description of the world at the time of any given Nuham's birth, they can *perfectly predict* exactly what that Nuham will be doing at any moment. So, if they had a description of the world at the time of a Nuham's birth, they could perfectly predict everything that Nuham would do throughout his or her life.

Subjects in both conditions were then asked three questions.

The first question essentially asked whether the recently discovered interesting fact, which had been stated to be true of all Nuhams, is true of a particular Nuham about to choose what to order for dinner.[1] This question was designed to tell whether or not a subject was paying attention. If a subject answered it incorrectly, his or her other data were excluded from all further analyses. A strong majority of subjects in both conditions answered correctly. The second question asked whether the recently discovered interesting fact is true of humans.[2]

[1] For example, the first question in the predictability condition:

> Imagine a Nuham named Nicole. Nicole is trying to choose what to order for dinner at a restaurant.
>
> If Nuham scientists had a description of the world at the time of Nicole's birth, could they perfectly predict what Nicole would choose to order?

Subjects could respond to this question by marking a box labeled "no" or a box labeled "yes".

[2] For example, the second question in the causation condition:

It was designed to make subjects think harder than they might otherwise about the fact, but also to really investigate subjects' beliefs about causation and predictability. A strong majority in both conditions indicated that the recently discovered interesting fact about Nuhams is *not* a fact about humans; most subjects seem to be indeterminists.[3]

The third question asked whether a particular Nuham act could have been performed with a free will. This question was the primary focus of the study. Subjects in the causation condition were told about Jonathan:

> Imagine that when a Nuham named Jonathan turns 30, he embezzles a large sum of money. This act was completely caused by things that happened before it, which were themselves completely caused by things that happened before them ... and so on back to things that happened before Jonathan was born.

Subjects in the predictability condition were also told about Jonathan:

> Imagine that when a Nuham named Jonathan turns 30, he embezzles a large sum of money. When Jonathan was born, Nuham scientists used a description of the world at the time of Jonathan's birth to predict that he would embezzle the money in exactly the way he does 30 years later.

Subjects in both conditions were then asked to indicate their agreement or disagreement with the statement

> It could be that Jonathan acted with a free will when he embezzled the money.

> Now think about our own universe and about the humans in it. Do you think that this fact that is true of the Nuhams and their universe is also true of us and our universe? That is, do you think that everything any given human does is completely caused by things that happened before, which are themselves completely caused by things that happened before them ... and so on back to things which happened before that human was born?

Again, subjects could respond to this question by marking a box labeled "no" or a box labeled "yes".

[3]Interestingly, *more* subjects in the predictability condition indicated that humans are not like Nuhams. 79% of subjects asked indicate that not all human actions are completely caused, but 92% indicate that not all human actions are perfectly predictable. This trend was not statistically significant, but a similar trend in a pilot study with a larger subject pool *was* significant.

To respond, they marked one of seven boxes going from "strongly disagree" to "neither agree nor disagree" to "strongly agree."

2.2 Primary results

Responses to this third question differed significantly[4] between the two conditions, with subjects in the causation condition tending to deny free will more than subjects in the predictability condition. The responses were scored on a scale from 1, corresponding to strong *disagreement* with the possibility of free will, to 7, corresponding to strong *agreement* with the possibility of free will. The mean response in the causation condition was 3.38, with 30% of responses on the "agree" side of neutral – a strong trend towards incompatibilism. The mean response in the prediction condition was 4.56, with 63% of responses on the "agree" side of neutral – a strong trend towards compatibilism.

So my hypothesis was confirmed. Not only was the difference in responses in the right direction, but the average responses in the two conditions were on opposite sides of a neutral response. Subjects in the causation condition tended to give incompatibilist responses, and subjects in the predictability condition tended to give compatibilist responses. Other than the difference between complete causation and perfect predictability, there seem to be no differences between the vignettes in the two conditions that could explain this difference in responses. I am strongly inclined to conclude from this study that ordinary people see their concepts of causation and free will as significantly less compatible than their concepts of predictability and free will.

3 Implications

So the simple difference between predictability and causation does make a difference in ordinary judgments about free will. This difference does not completely explain the results

[4] $t(47) = 2.25$, $p = 0.029$

of *all* previous studies of ordinary judgments about free will, though it might partially explain the results of *some* studies, including the Jeremy Hall study by Nahmias et al. and the Universe A study by Nichols and Knobe. More than that, this difference is an interesting component of the ordinary approach to free will. But what does this difference really show us? Is there perhaps, according to ordinary people's implicit beliefs, some aspect of causation the negation of which is necessary for free will? Further research is needed, but the results of this study do offer some indications of just what ordinary people take to conflict with free will.

The results from the predictability condition seem to indicate that ordinary people are compatibilists about determinism and free will, while the results from the causation condition might seem to indicate the opposite. It seems safe to assume that subjects' notions of free will do not differ between the two conditions, and it also seems safe to assume that subjects normally ascribe free will to beings capable of embezzling money, only denying free will when they take some abnormal aspect of the situation to conflict with it. We should then conclude that subjects infer something about the imaginary universe in the causation condition – and not in the predictability condition – that they take to conflict with free will.

And though subjects in the causation condition must take *something* to conflict with free will, they cannot take *determinism*, plain and simple, to conflict with free will, for subjects in the predictability condition do not take anything to conflict with free will, and it seems they have little choice but to infer determinism.[5] If an action can be perfectly predicted based on a description of the world at a certain time, that description of the world – plus additional general knowledge – allows the predictor to correctly assign 100% probability to that action,

[5]One alternative explanation of subjects' responses in the predictability condition is the following: many subjects believe in a deity with perfect knowledge of future events and have reconciled this belief with the belief that humans often act with free will. These subjects might, in the same way, reconcile a belief that Nuham actions are perfectly predictable with a belief that Jonathan acts with free will.

This explanation requires subjects to make an analogy between the Nuham scientists' powers of prediction and God's foreknowledge. It seems unlikely that subjects would do so, since they are told that Nuham scientists make predictions "[b]ased on a description of the world" at a certain time. This should lead subjects to conceive of the Nuham scientists' powers of prediction very differently from the way they probably conceive of God's foreknowledge.

leaving a 0% probability for every other possible outcome. This is surely a sufficient condition for the action's being determined. It is just as surely what subjects inferred about the action in the predictability condition, since predictability is a straightforward notion about which subjects are unlikely to be confused.

Causation, though, is not a very straightforward notion, even for the metaphysicians who study it professionally. It seems plausible, then, that subjects' notions of causation carry with them something over and above what is necessary for perfect predictability or for determinism. Indeed, since complete causation is the only thing subjects can infer directly from the text in the causation condition that they cannot infer in the predictability condition, it must be something about or implied by *causation itself* that conflicts with free will in the causation condition but not in the predictability condition. We can now do little more than speculate about the precise nature of this aspect of the ordinary notion of causation that conflicts with free will and is not obviously implied by perfect predictability, but it is an area open to much further research.

Because the results from the predictability condition are fairly straightforward, this study, like those by Nahmias et al., indicates that ordinary people are really compatibilists about free will and determinism. In addition, this study offers a potential explanation of apparent ordinary incompatibilists: those people who seem to see a conflict between free will and determinism might actually be inferring complete causation from determinism and then seeing a conflict between free will and causation. This is a plausible scenario, since complete causation might easily be seen as the best explanation of perfect predictability. But the more plausible this scenario is, the more surprising the results of this study should be. If philosophers see complete causation as the best and most obvious explanation of perfect predictability, we are at odds with the ordinary people, who seem to be unwilling or unable to infer causation from predictability.

Previous work investigating ordinary people's intuitions about free will and determinism has treated complete causation and perfect predictability as interchangeable standards for

determinism. This might have been a mistake, as ordinary people treat complete causation and perfect predictability as distinct entities: this study indicates that they refuse to infer the former from the latter. What's more, this study indicates that ordinary people differentiate strongly between complete causation and perfect predictability when making judgments about free will. Just why they differentiate as they do is worth investigating further, both empirically and from armchairs. However, regardless of *how* or *why* ordinary people make this distinction, any further investigation of ordinary people's intuitions about free will and determinism should take into account *that* ordinary people *do* make this distinction: that they see complete causation as *incompatible* with free will and perfect predictability as *compatible* with it.

References

Nahmias, E. (2006). Folk fears about freedom and responsibility: Determinism vs. reductionism. *Journal of Cognition and Culture*, 6:215–237.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research*, 73:28–53.

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