Experimental Philosophy and the Compatibility of Free Will and Determinism : A Survey

Cova, Florian; Kitano, Yasuko

How to cite

COVA, Florian, KITANO, Yasuko. Experimental Philosophy and the Compatibility of Free Will and Determinism : A Survey. In: Annals of the Japan Association for Philosophy of Science, 2014, vol. 22, p. 17–37. doi: 10.4288/jafpos.22.0_17

This publication URL: https://archive-ouverte.unige.ch/unige:109413
Publication DOI: 10.4288/jafpos.22.0_17

Annals of the Japan Association for Philosophy of Science Vol.22 (2014) 17~37

Experimental Philosophy and the Compatibility of Free Will and Determinism: A Survey

Florian COVA* and Yasuko KITANO**

Abstract

The debate over whether free will and determinism are compatible is controversial, and produces wide scholarly discussion. This paper argues that recent studies in experimental philosophy suggest that people are in fact "natural compatibilists". To support this claim, it surveys the experimental literature bearing directly (section 1) or indirectly (section 2) upon this issue, before pointing to three possible limitations of this claim (section 3). However, notwithstanding these limitations, the investigation concludes that the existing empirical evidence seems to support the view that most people have compatibilist intuitions.

Key words: experimental philosophy; free will; moral responsibility; determinism

Introduction

The recently developed field of experimental philosophy advocates a methodological shift toward the use of an experimental methodology, in order to make progress on problems in philosophy (Knobe and Nichols, 2008). On a *narrow* view, experimental philosophy is an experimental investigation of our intuitions about philosophical issues; on a *larger* view, it encompasses all tentative attempts to make progress on philosophical questions using experimental methods (Cova, 2012).

Recently, a growing number of studies in experimental philosophy are addressing the relationship between free will and moral responsibility. Most of these aim (i) at testing whether people have the intuition that an agent in a deterministic universe cannot have free will and be morally responsible for his or her actions by (ii) probing their intuitions on small vignettes[1]. In this paper, we address (i), i.e. the *compatibility question*: whether people really have incompatibilist intuitions on the compatibility of free will with determinism[2].

* Swiss Centre for Affective Sciences, University of Geneva
** Graduate School of Arts and Sciences, The University of Tokyo

[1] Sommers (2010) criticizes both (i) and (ii).

[2] However, we shall not go into methodological discussions about (ii) in this paper.

Whether they do is an important matter, in terms of both theory and practice. In regards to practice, Greene and Cohen (2004) argue that neurosciences, because they promote a deterministic view of human decision making, they will lead people to abandon their beliefs in free will and moral responsibility, and thus to abandon a retribution-based conception of legal punishment in exchange for a more consequentialist perspective. However, this prediction rests on the premise that people are "natural incompatibilists", i.e. that they pre-theoretically and in absent of any philosophical training will consider free will to be incompatible with determinism. If it were shown that, on the contrary, people pre-theoretically believe free will to be compatible with determinism and, thus, were "natural compatibilists", this whole argument would collapse[3].

More broadly, these experimental findings can impact the perception of the free will problem, particularly outside of the circle of free will specialists. Non-specialists could assume that it is obvious that determinism and free will are incompatible and thus fail to appreciate specialists' efforts to solve this problem. For example, scientists like Sam Harris (2012) and Jerry Coyne (on his blog) claim that current scientific evidence show that we have no free will. But their arguments rely on a very specific incompatibilist conception of free will and moral responsibility, which they nonetheless assume represents the "common view" of society. In this situation, showing that people do not unanimously believe that free will requires indeterministic choice would contribute to bridging the gap between scientists and philosophers, as it would demonstrate to the former that questions regarding the common conception of free will are far from obvious, and thus still worth investigating.

In terms of theory, some have argued that folk intuitions give a dialectical advantage to the philosophical position that is closer to common sense, and help determine who has to carry the burden of proof (e.g. Nahmias et al. 2006). However, we will not explicitly address such meta-philosophical issues. Instead, we remain focused upon the possibility of synthesizing the available experimental results in order to give an answer to the purely descriptive question known as the *compatibility question.* After surveying experiments bearing directly (section 1) or indirectly (section 2) on this issue[4], we will conclude that current results support the conclusion that people

We'd like to thank an anonymous referee for pointing out that readers who are not familiar with experimental approaches in philosophy might have trouble identifying the precise question we address in this paper.

[3] See Cova (2011).

[4] The present paper intends to be a survey of the experimental studies bearing on the compatibility question, some of which mainly focus upon moral responsibility but yet have implication to free will ascription, and not a meta-analysis of those (an impossible exercise, since these studies differ widely in methods and objectives). For a meta-analysis of some of the studies presented in this paper, see Feltz and Cova (submitted).

are "natural compatibilists", but also address three possible limitations to this claim (section 3).

1. Folk intuitions about the compatibility of free will and determinism

The 'Natural Compatibilist' Hypothesis

In one of the first studies in experimental philosophy of free will, Nahmias and his colleagues (2005, 2006) gave to 39[5] participants the following story that describes an agent living in a world ruled by a Laplacean determinism (i.e. in a world in which each state of the world could be deduced from the conjunction of the complete description of the world at an anterior state and the laws of nature):

> *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. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25, 2150 AD, 20 years before Jeremy Hall is born. The computer then deduces from this information and the laws of nature that Jeremy will definitely rob Fidelity Bank at 6:00 pm on January 26, 2195. As always, the supercomputer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 pm on January 26, 2195.*

Some participants were then asked:

> *Imagine such a supercomputer actually did exist and actually could predict the future, including Jeremy's robbing the bank (and assume Jeremy does not know about the prediction):*
> *Do you think that, when Jeremy robs the bank, he acts of his own free will?*

While others were asked:

> *Do you think that when Jeremy robs the bank, he's morally blameworthy for it?*

To the first question, 76% of 21 participants answered that Jeremy acted of his own free will. To the second, 83% of 18 participants answered that Jeremy was morally

[5] When possible, we report the sample size, since it is an important data to judge the generalizability of a study's results. However, this is not always possible. Overall, sample size have tended to increase with time, as experimental philosophy became more methodologically demanding. Participants in these studies are generally either undergraduate students or participants recruited online (for example, through Amazon Mechanical Turk), though they are notable exceptions, that we will point out.

blameworthy. In another version Jeremy did not rob a bank but saved a child from a burning building. To this scenario, 68% of 22 participants answered that Jeremy saved the child of his own free will and 88% of 18 participants judged that Jeremy was praiseworthy for having saved the child. Finally, in a third version of the story, Jeremy decided to go jogging. In this case, 79% of 19 participants answered that Jeremy went jogging of his own free will.

In addition to this version (the SUPERCOMPUTER case), they have obtained similar results with two other kinds of scenarios. First, they gave 69 participants scenarios about two IDENTICAL TWINS who were supposed to live in a world in which one's beliefs and values are caused completely by the combination of her or his genes and the environment she or he is in. As they have been separated at birth, the first twin (Fred) is raised in family that teaches to value money above all else while the second (Barney) is raised in a family that teaches him to value honesty above all. Then the following happens:

> *One day Fred and Barney each happen to find a wallet containing $1000 and the identification of the owner (neither man knows the owner). Each man is sure there is nobody else around. After deliberation, Fred Jerkson, because of his beliefs and values, keeps the money. After deliberation, Barney Kinderson, because of his beliefs and values, returns the wallet to its owner. Given that, in this world, one's genes and environment completely cause one's beliefs and values, it is true that if Fred had been adopted by the Kindersons, he would have had the beliefs and values that would have caused him to return the wallet; and if Barney had been adopted by the Jerksons, he would have had the beliefs and values that would have caused him to keep the wallet.*

Having read this scenario, 76% of 34 participants judged both that Fred kept the wallet of his own free will and Barney returned it of his own free will. Moreover, 60% of 35 participants judged that Fred was blameworthy and 64% judged that Barney was praiseworthy.

Finally, a third kind of scenarios described a RE-CREATING UNIVERSE that is identically recreated again and again so that everything must take place the exact same way. In this case the participants were asked both to judge whether Jill (a person living in this world) decided to steal a necklace of her own free will and whether "it would be fair to hold her morally responsible (that is, blame her) for her decision to steal the necklace." Most participants offered consistent judgments; overall, 66% judged that Jill acted of her own free will, and 77% judged her to be morally responsible.

Taken together, such results strongly suggest that people tend not to consider free will and determinism to be incompatible. Let's call this possibility the "Natural Compatibilist Hypothesis" (hereafter NCH): people are, mostly, "natural compati-

bilists". Though NCH explain the data we just surveyed in a straightforward way, the data we present in the following section are not so easily accommodated.

Framing effects and conflicting intuitions

The trouble for NCH comes from two experiments by Nichols and Knobe (2007; see also Nichols, 2006). In both experiments, participants were presented with the following description of two different universes:

> *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.*
>
> *Now imagine a universe (Universe B) in which almost everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French Fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it did* ***not*** *have to happen that Mary would decide to have French Fries. She could have decided to have something different.*
>
> *The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision— given the past, each decision has to happen the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision does not have to happen the way that it does.*

In the first experiment, participants were then asked which one of these two universes was more like ours. Nearly all participants (90%) answered 'Universe B'. Then, participants in the CONCRETE CONDITION received the following scenario:

> *In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and 3 children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family.*
> *Is Billy fully morally responsible for killing his wife and children?*

In this condition, most subjects (72%) gave the compatibilist answer according to

which the agent was fully morally responsible. These results are consistent with those obtained by Nahmias and his colleagues. But, let's consider now the ABSTRACT CONDITION. Participants in this condition had no scenario to read but just received the following question:

> *In Universe A, is it possible for a person to be morally responsible for their actions?*

In this condition, most subjects (86%) gave the incompatibilist answer. How is it possible that participants' answers are so inconsistent? A first solution could be that the concrete condition scenario is too long and complex for participants to keep track of the relevant fact. Nichols and Knobe gave to participants a shorter version of the CONCRETE CONDITION: 50% gave the compatibilist answer, which is still higher than in the ABSTRACT CONDITION. Thus, another solution must be advanced.

Three accounts of conflicting intuitions

Many accounts of this phenomenon have been advanced. Three have been at the center of discussions and subject to experimental testing[6]:

(1) The 'Abstract vs. Concrete' hypothesis

According to Sinnott-Armstrong (2008), differences in participants' answer could stem from the fact that processing abstract and concrete stimuli engage different cognitive mechanisms. Drawing on the Construal Level Theory, a psychological theory according to which distant phenomena (along any of these four dimensions: spatial/temporal/social/hypothetical) are construed at a higher, more abstract level, Weigel (2011) asked 142 participants to read the following scenario:

> Scientists have discovered that everything that happens is completely caused by whatever happened before it. This is and has always been true: all events, including human choices and decisions, happen because of something that completely caused it. For example, when I ate French fries for lunch, my decision caused me to eat the French Fries, but the decision was also completely caused by what happened before it. Prior events caused my decision to eat the French Fries, and my decision determined what I ate. Keep that in mind as I tell you what happened with Bill.
>
> Bill became attracted to his secretary, and he decided that the only way to be with her would be to kill his wife and three children. Knowing that it would be impossible to escape from his house in the event of a fire, he set up a device in

[6] For a fourth account that has not been discussed so far, see Mandelbaum and Ripley (2012).

Table 1 Results from Nichols & Knobe (2007)

	Agent in indeterministic universe	Agent in deterministic universe
High Affect case	95%	64%
Low Affect case	89%	23%

his basement that burned down the house and killed his family.

In the NEAR CONDITION, before reading the scenario subjects were asked to imagine that they were going to hear the lecture *in a few days* and then to predict how they would answer the questions *in a few days*. In the DISTANT CONDITION, "in a few days" was replaced by *in a few years*. Weigel found that participants were more likely to give incompatibilist answers (i.e. to answer that Bill did not make his decision freely) in the DISTANT CONDITION — a likely explanation being that distance led participants to think in a more abstract way and that abstract thought promotes incompatibilist intuitions.

(2) The 'Performance Error Model'

To explain their own results, Nichols and Knobe (2007) advance what they called the 'Performance Error Model' (hereafter, PEM): it might be that compatibilist answers are emotionally driven and that people are more compatibilist in the concrete case because the situation described (a murder) is emotionally loaded. To test this hypothesis, Nichols and Knobe designed two new conditions. The LOW AFFECT condition was the following:

> *As he has done many times in the past, Mark arranges to cheat on his taxes. Is it possible that Mark is fully morally responsible for cheating on his taxes?*

While the HIGH AFFECT condition was the following:

> *As he has done many times in the past, Bill stalks and rapes a stranger. Is it possible that Bill is fully morally responsible for raping the stranger?*

In each condition, for half of the subjects, the question stipulated that the agent was in (deterministic) Universe A while, for the other half, the question stipulated that the agent was in (indeterministic) Universe B. Table 1 describes, for each combination, the proportion of participants who answered 'yes' to whether the agent could be fully morally responsible for his action.

Contrary to what NCH would have predicted, and as predicted by PEM, participants in the LOW AFFECT condition tended to judge that the agent situated in a deterministic universe could not be responsible for cheating on his taxes. For Nichols and Knobe, these results support PEM over NCH by showing that participants have

compatibilist intuitions when affective reactions are kept low enough not to bias their judgments. Additionally, these results also pose a problem for the 'Abstract vs. Concrete' hypothesis, since both the 'high affect' and 'low affect' conditions seems to be concrete cases.

However, PEM faces a number of problems. First, this account is incompatible with the fact that, in Nahmias' SUPERCOMPUTER (JOGGING) case, people were compatibilists about Jeremy going jogging, while this is clearly a low affect case.

Second, Nichols and Knobe's results have not been fully replicated. If their results for the ABSTRACT CONDITION have been replicated cross-culturally[7], the difference between the HIGH AFFECT and the LOW AFFECT cases has not. The only published paper to directly attempt at such a replication failed twice and found both times that participants gave mostly incompatibilist answers in both cases (Feltz *et al.*, 2009). Thus, it seems that PEM does not rest on a firm ground[8].

Finally, in a recent study, Cova, Bertoux and their colleagues (2012) have given Nichols and Knobe's CONCRETE CONDITION and Nahmias *et al.*'s SUPERCOMPUTER (BANK) case to 12 patients suffering from a behavioural variant of frontotemporal dementia, a neurodegenerative disease accompanied by a deficit in emotional responses. Contrary to what PEM would have predicted given their lack of emotional reactions, these patients were no more incompatibilist than control participants and gave mostly compatibilist answers[9].

Thus, it seems that there is a certain number of empirical data PEM cannot account for. Let's however note that the first problem—the fact that Nahmias and his colleagues found mostly compatibilist answers for a neutral case in the SUPERCOMPUTER (JOGGING) case—can be addressed by adding a supplementary hypothesis: the JOGGING scenario is set in our actual world, while Nichols and Knobe's cases are set in imaginary universes. So, it might be that this difference in design can account for the difference in results. And indeed, Roskies and Nichols (2008) found that, *ceteris paribus*, people were more prone to judge agents morally responsible for their actions when the scene took place in our world rather than in an alternate universe. However, this auxiliary hypothesis does not solve *all* the problems PEM has to face.

(3) An Error-Theory for Incompatibilist Intuitions

How then are we to reconcile these conflicting results? Nahmias and Murray

[7] See Sarkissian *et al.* (2010). The study gathered 231 undergraduate students from four countries: United States, India, Hong Kong, and Columbia.

[8] A recent meta-anaysis of 30 published and unpublished studies shows that if affect has an effect on ascriptions of free will, this effect is far too small to explain the difference between the abstract and concrete cases (Feltz and Cova, submitted).

[9] Control groups were 10 healthy participants and 10 patients suffering from Alzheimer's disease.

(2010; see also Murray and Nahmias, in press) have recently proposed a solution. But to understand this solution, we must first describe another puzzling pair of cases. Nahmias (2006) used a pair of scenarios describing two different kinds of determinism. The first scenario (PSYCHOLOGICAL DETERMINISM) described a planet similar to ours (Erta) inhabited by people called the Ertans. On this planet, Ertan psychologist have discovered that the Ertan's thoughts, desires, and plans occurring in her or his mind completely cause all the decision the Ertan makes. The psychologists also have discovered that these thoughts, desires and plans are completely caused by the Ertan's current situation and the antecedent events she or he has been through. In the second scenario (NEUROLOGICAL DETERMINISM), the same planet was described but, this time, the neuroscientists have discovered that the decision the Ertan makes is completely caused by the specific neural processes occurring in his or her brain; these neural processes are completely caused by the Ertan's current and the antecedent events she or he has been through. For both scenarios, participants were asked if Ertans could act of their own free will and whether they deserved to be given credit or blame for their actions. In the PSYCHOLOGICAL DETERMINISM condition, 72% of 25 participants answered positively to the first question and 77% of 22 participants answered positively to the second question. In the NEUROLOGICAL DETERMINISM condition, only 18% of 22 participants answered positively to the first question and 19% of 21 to the second question[10]. Otherwise said: people give mostly compatibilist answers in the first case and mostly incompatibilist answers in the second case.

How are we going to make sense of these results? Nahmias suggests that this asymmetry arises because people take neurological determinism, contrary to psychological determinism, to imply that people's mental states do not have a role to play in the generation of their action and are 'bypassed'. This is, he argues, why the presence of determinism in the PSYCHOLOGICAL DETERMINISM case does not prevent participants of giving mostly compatibilist answers. In the NEUROLOGICAL DETERMINISM case, though, the neurological description leads them to believe that the agents' mental states do not play any role in the generation of their actions. In this case, it seems natural to withhold attributions of free will and moral responsibility.

Applying this explanation to Nichols and Knobe's cases, Nahmias and Murray made the following predictions:

i. *Ceteris paribus*, Nichols and Knobe's description of determinism leads to more confusion of determinism (the thesis that every human action is fully caused by prior events) with 'bypassing' (the thesis that human beings' mental states do not play a causal role in the generation of human action and are thus 'bypassed') than the three descriptions used by Nahmias *et al.* (hence the greater number

[10] These results were replicated on a larger sample (1,124 undergraduate students) in Nahmias *et al.* (2007).

of incompatibilist answers in Nichols and Knobe's study).

ii. People are less likely to conflate determinism with 'bypassing' in concrete than in abstract cases (hence the difference between the two kinds of cases). This difference might be due to the fact that the concrete descriptions of agents and actions prime people to think about the effectiveness of agent's mental states.

To test these predictions, Nahmias and Murray gave to 249 participants a scenario among a set constituted by Nichols and Knobe's ABSTRACT CONDITION and CONCRETE CONDITION (in Universe A) as well as an abstract and a concrete version of the RE-CREATING UNIVERSE case. Participants were not only asked if agents in these scenarios deserved praise or blame and acted from their own free will, but they were also asked questions designed to probe their understanding of determinism.

Nahmias and Murray's results matched their predictions. First, they found that compatibilist intuitions were highly correlated to a good understanding of determinism (that is: an understanding of determinism that doesn't conflate it with "bypassing") and that people who gave incompatibilist answer were far more susceptible to believe that the determinism entailed "bypassing". Second, they found that Nichols and Knobe's description of Universe A scenario led a great number of participants to think that mental states were "bypassed" than Nahmias *et al*'s description of the RE-CREATING UNIVERSE. Finally, they observed that people in the "abstract" condition were more prone to adopt the "bypassing" interpretation of determinism. They conclude that most incompatibilist answers are only apparent incompatibilist answers, because most of them were the product of a bad understanding of determinism.

This Error-Theory of incompatibilist answers (ET) seems to be currently the best explanation available. It explains (i) the difference between Nichols and Knobe's and Nahmias *et al.*'s results and (ii) the difference between abstract and concrete cases. It can also explain (iii) why people are more likely to give incompatibilist answers when the scenario is set in our actual world: since (most) people do not believe that our world is a world in which our mental states are powerless, they are less likely to take determinism as entailing 'bypassing'.

However, De Brigard and his colleagues (2009) made the following objection to ET: having systematically varied whether the agent's action were produced by psychological or neurological causes, they never found a difference between those two conditions. For example, in their second study (on 60 undergraduate students), the PSYCHOLOGICAL CONDITION was set in the following way:

> Dennis and John have been friends for thirty years who always meet for a weekly walk. Dennis has been away on vacation for a month and so the friends have not been able to go on their walk until last week. On their walk last week they passed a jogger on their normal trail. Seemingly unprovoked, Dennis ran up to

> the jogger and punched him in the stomach multiple times. Shortly after this incident Dennis was diagnosed with a **psychological** illness that causes him to manifest uncontrollably aggressive behavior which in turn caused him to hit the jogger.
>
> On a scale of 1–7 how responsible is Dennis for hitting the jogger?

In the NEUROLOGICAL CONDITION, the word '**psychological**' was replaced by the word '**neurological**'. For both condition, the mean answer was 3.8. Thus, there were no difference.

However, in De Brigard *et al.*'s case, the psychological state is an illness and is stated as "uncontrollable": it is then likely that people will tend to consider this kind a psychological state as no more part of the agent's true desires than a neurological state (both explanation are "mechanistic" explanations, rather than explanations in terms of reasons). Thus, we don't think that these experiments provide a counter-example to ET, as long as we understand Nahmias as saying that stemming from the agent's mental states is necessary but clearly not sufficient for moral responsibility.

Still, one might argue that we miss the point: not only De Brigard *et al.* did not find a difference between the PSYCHOLOGICAL CONDITION and the NEUROLOGICAL CONDITION — they also found high attributions of responsibility in both cases. How can it be if people consider the psychological and neurological illness as alien to the agent's true motivations? Our answer will be that most people can still hold the agent's responsible for not resisting this alien impulse — i.e. for a lack of control they attribute to the agent's true motivational states[11]. Thus, all the results obtained by De Brigard can be explained in the framework of ET by making the two following reasonable assumptions: (i) people consider that certain mental states (as illnesses) are not really part of the agent's motivation and (ii) even when agents are driven by alien sources, they can still be judged responsible if they are perceived as failing to refrain this impulse when they could.

Thus we conclude this section by claiming that ET is the best currently available. As suggested by our discussion, it is not a full theory of the folk conception of free will and moral responsibility and must be refined to account for all the data (rather than being just compatible with them), but it fully fulfills its aim to account for folk

[11] Even if the impulse is described as "uncontrollable", one should keep in mind that participants constantly add to vignettes assumptions drawn from their background beliefs. For example, it could be that participants refuse to imagine that one cannot refrain to hit a person. Also, there is an ambiguity in the expression "uncontrollably aggressive behavior": one could understand this expression as meaning that this behavior is uncontrollable *once triggered*, but that it is still up to the agent to prevent the manifestation of this behavior. For example, if I say that, for Dennis, drinking alcohol causes uncontrollably aggressive behavior, it does not mean that Dennis cannot prevent this aggressive behavior (for example, by avoiding alcohol).

intuitions about the compatibility of free will and determinism.

2. Testing the premises of the philosophical arguments for incompatibilism

The experiments we surveyed so far tested folk intuition about the compatibility question. However, most philosophical arguments for the incompatibility of free will with determinism do not use this kind of direct intuitions. Instead, most arguments start from intuitive premises they show to ultimately conflict with the compatibility of free will with determinism. Thus, it could be that common sense is incompatibilist in the sense that folk intuitions about these premises ultimately conflict with their direct intuitions about the compatibility of free will with determinism. This is why, in this section, we survey experiments on folk intuitions about such premises.

The Principle of Alternate Possibilities

One of the most famous principles used in support of incompatibilism is the Principle of Alternate Possibilities (PAP), according to which no one is responsible of what one has done if one could not have done otherwise. On a certain interpretation of "could have done otherwise", according to which determinism is incompatible with the ability to do otherwise, endorsing this principle leads quite directly to endorse incompatibilism about free will and moral responsibility. Nevertheless, there are two ways of rebutting this kind of arguments in favor of incompatibilism:

- *The interpretation question*: one can argue that the meaning of "could have done otherwise" that is relevant for free will and/or moral responsibility is perfectly compatible with determinism.
- *The truth question*: one can also argue that the PAP is just false. The most famous arguments for the falsity of PAP are the Frankfurt cases, in which we are supposed to have the intuition that an agent is morally responsible for his action even if he could not have done otherwise.

There are a few data relevant to the interpretation question. In their experiments on the SUPERCOMPUTER cases, Nahmias and his colleagues (2005) also asked participants whether the agent could have chosen *not* to act the way he did. Results varied along the nature of the action: for the negative action (robbing the bank), 67% of 21 participants answered positively, but only 38% out of 21 in the positive action condition (saving the child), and only 43% out of 14 in the neutral action condition (going jogging). Also, for the IDENTICAL TWINS case, they asked whether the twins could have done otherwise, and 76% of 31 participants answered that both Fred and Barney could have done otherwise. These results indicate that there is a significant proportion of participants willing to attribute the ability to choose and do

otherwise to agents in a deterministic universe (at least 38%), but there is too much variation from one case to the other to draw support either the compatibilist or the incompatibilist interpretation of "could have done otherwise", and we cannot know whether participants are interpreting this expression in the sense relevant for free will and moral responsibility.

Other relevant results can be found in Nahmias and his colleagues (2004) on the phenomenology of free will. They gave participants the following survey:

> Imagine you've made a tough decision between two alternatives. You've chosen one of them and you think to yourself, 'I could have chosen otherwise' (it may help if you can remember a particular example of such a decision you've recently made). Which of these statements best describes what you have in mind when you think, 'I could have chosen otherwise'?
>
> A. 'I could have chosen to do otherwise even if everything at the moment of choice had been exactly the same'.
>
> B. 'I could have chosen to do otherwise only if something had been different (for instance, different considerations had come to mind as I deliberated or I had experienced different desires at the time)'.
>
> C. Neither of the above describes what I mean.

35% of 96 participants chose answer A, 62% answer B and 3% answer C. This suggests that more than half of participants (62%) endorsed a compatibilist interpretation of 'could have chose otherwise' (that is: an interpretation according to which an agent subject to determinism could still choose otherwise). However, the problem is that it is hard to determine whether participants are really expressing the meaning of 'could have chosen otherwise' they consider relevant for free will and moral responsibility.

About the truth question, more straightforward data can be found. Woolfolk and his colleagues (2006) presented to 48 participants a case in which a man, Bill, is compelled by a 'compliance drug' to kill one of his friends. There were two conditions: in one case, Bill already desired (and had planned) to kill his friend (because this friend had an affair with his wife; thus Bill 'identified' himself with his action) while in the other case Bill did not want to kill his friend (so, he didn't 'identify' himself with his action, where 'identifying' oneself with one's action means that this action is in accordance with one's deepest desires and values). Participants were asked on a 7-points scale whether Bill was responsible for his action and whether he could have done otherwise. Participants were more likely to judge that Bill was responsible for having killed his friend in the first case (3.25 *vs.* 2.25), even though they think that he *was not free to do other than he did* in both cases. These results suggest that moral responsibility can be independent from the ability to do otherwise (though this particular study provides no data about participants' ascriptions of free will).

A more direct confirmation of Frankfurt's intuitions can be found in Miller and

Feltz (2011), who used 'Frankfurt-style cases', as for example, a case in which Mr. Jones decides on his own to steal a car but would have been forced to do it by an evil neurosurgeon if he had not. In this case, 52 participants were asked whether Mr. Jones could have done anything other than decide to steal the car. Among those who answered 'no' (and thus passed the comprehension check), a great majority answered that Mr. Jones was morally responsible for stealing the car and deserved blame (participants had to rate the agent's moral responsibility and blameworthiness on 7-points scales, and the average scores were respectively 5.27 and 5.40)[12]. Taken together, all these results suggest (i) that most people consider that the possibility to choose or do otherwise is not necessary for moral responsibility and (ii) that most people consider that there is a sense in which an agent subject to determinism can do or choose otherwise.

The Ultimacy Argument

Another argument for incompatibilism is to claim that our shared concept of free will requires power and abilities which are not compatible with determinism, as the ability to be the ultimate source of one's actions (e.g. Strawson, 1994). This is classically called the Ultimacy Argument. However, available empirical evidence does not seem to support this argument.

Monroe and Malle (2010) have asked 180 people to define what they meant by "free will" — the exact question being: "Please explain in a few lines what you think it means to have free will" (p.214). Their analysis of participants' answers reveal that respondents didn't commit themselves to an incompatibilist understanding of "free will" in their definitions: most of them defined it as "the ability to make a choice" (65%), "doing what you want" (33%) and "acting without external or internal constraints" (29%) — all conceptions that are *prima facie* compatible with determinism and doesn't appeal to classic libertarian features. Monroe and Malle thus conclude

[12] For a replication, see Cova (forthcoming). As pointed out by a reviewer for this journal, the existing literature does not allow us to draw a similar conclusion for free will, due to a lack of data. Further researches should investigate this question. As a start, we ran our own study. 40 participants living in United States and recruited through Amazon Mechanical Turk (for a salary of $0.3) received Miller and Feltz's vignettes then were asked two comprehension checks ("Was it possible for Mr. Jones to avoid stealing the car at 12:00am on October 7?", "At 12:00am on October 7, could Mr. Jones have done anything other than steal the car just then?") and three questions ("Did Mr. Jones freely steal the car?", "Did Mr. Jones steal the car on his own free will?", "Is Mr. Jones morally responsible for stealing the car?"). Among the 32 participants who answer "NO" to both comprehension checks, 94% answered "YES" to the first question, 91% answered "YES" to the second question, and 91% answered "YES" to the third question. These results give us *prima facie* reasons to think that the conclusion we draw in this section extend to participants' ascriptions of free will, and that most people think that free will do not require the freedom to do otherwise.

that "the social-linguistic community as a whole appears to define free will as a choice that follow one's desire and is not internally or externally constrained" (p.215).

Inspired by Monroe and Malle's study, Stillman (2011) and his colleagues have studied autobiographical accounts of free and unfree actions. They asked 99 participants to write about their own past behaviors that they deemed to be free or, on the contrary, felt had not reflected their own free will. By comparing the two kinds of narratives along a certain number of predefined dimension, Stillman and his colleagues have found that "participants in the free will condition described events in which they had acted against external forces, achieved goals, evinced conscious thoughtfulness, had a positive outcome, and behaved consistently with their morals. Participants in the free condition were also more likely than those in the unfree condition to report acting in a manner consistent with their enlightened self-interest" (p.388).

Thus, these studies stress the importance of choice and of the ability to act according to what one wants without being stopped by external constraints. However, these features are those on which compatibilist typically insist. On the contrary, we must note that the traditional incompatibilist requisites were never brought forward in Monroe and Malle's study.

The Manipulation Argument

Another argument for incompatibilism is the Manipulation Argument (see Pereboom, 1995 for an example). The Manipulation Argument is the name for a class of argument that goes like this:

i. Agents in manipulation cases have no free will.
ii. There is no relevant difference between agents in manipulation cases and agents living in a deterministic world.
iii. Therefore, agents living in a deterministic world have no free will.

Premise (i) is based on our intuitive reactions to manipulation cases: imagine that John is an ordinary man and that an evil neurosurgeon is able to implant in him the irrepressible desire to kill his wife. If John ends up killing his wife because of the neurosurgeon's manipulation, then it seems that we won't judge him morally responsible for killing his wife. However, the incompatibilists argue that there is no relevant difference between John and an non-manipulated agent living in a deterministic world: thus, if John is not free, nor are people living in a deterministic world.

However, it is possible to reject the argument by rejecting premise (ii): it might be that there are relevant differences between manipulated agents and agents living in a deterministic world. More precisely, it might be that we have the intuition that manipulated agents are unfree only because we consider them to lack cognitive or volitional capacities that people in deterministic universe are not supposed to lack.

Of course, incompatibilists who design the thought experiments underlying the Manipulation Argument consider that the manipulated agents in their cases lack such capacities, but this doesn't rule out the possibility that these cases trigger incompatibilist intuitions only to the extent that agents in these cases are perceived as lacking such capacities.

To test for this possibility, Sripada (2012) has given 240 participants one of two versions of a typical manipulation case. In the MANIPULATION version, Bill kills Mrs. White, but it turns out that Bill has been manipulated by the evil Dr. Z:

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

In the NO MANIPULATION version, manipulation is avoided, since Bill was not adopted by Dr. Z, though he still ends up killing Mrs. White.

In both conditions, participants had to rate on 7-points scale their agreement with statements about Bill's free will (e.g. "Bill killed Mrs. White of his own free will"), Bill's information about his action (e.g. "Bill killed Mrs. White based on false information about her, and he was deprived of any opportunity to learn the truth") and Bill's deep motivations (e.g. "Bill's killing Mrs. White does not reflect the kind of person who he truly is deep down inside")[13]. Sripada found that people were indeed less likely to attribute free will and moral responsibility to the agent in the MANIPULATION condition (4.09 vs 1.60 and 3.16 vs 1.49 respectively), but also more likely to perceive Bill as acting on the basis on corrupted information and in discordance with his deep self (4.43 vs. 6.07). Correlation and mediation analyses showed that free will attributions were significantly influenced by people's judgments about corrupted information and deep self discordance.

These results suggest that premise (ii) of the Manipulation Argument is problematic: manipulation cases trigger the intuition that a manipulated agent lacks free will only to the extent that the agent is perceived as lacking certain abilities that standard agents in deterministic universe are not supposed to lack. Moreover, they suggest that participants' intuitions are perfectly in accordance with compatibilism: it is because the agents lacks capacities typically stressed by compatibilist theories of free will that participants judge him as lacking free will. It is likely that if they did not perceive such impairments, they would be more likely to consider a manipulated agent as free. In a second study, Sripada tested this prediction by giving 120 participants a manipulation case in which he insisted on the fact that the agent's capacities were left untouched. He indeed found that participants were more likely

[13] Strangely, on these scales, a lower score meant a higher agreement

to consider the manipulated agent as free in this case than in the previous standard manipulation case (66% vs. 31%)[14].

Overall, then, it does not seem that NCH can be disputed on the ground that folk intuitions widely endorse a premise that could be shown to be incoherent with the compatibility of free will with determinism. Rather, it seems that the study we surveyed lend extra support to the thesis that folk intuitions are mostly compatibilist, by undermining the intuitiveness of incompatibilist premises.

3. The remaining questions

In the previous sections, we argued that, so far, empirical studies of folk intuitions about determinism, free will and moral responsibility favor the compatibilist side, by suggesting that laypeople tend to be naturally compatibilists. However, such a conclusion is too simplistic, for certain elements seem to complicate the picture by suggesting that certain factors can make folk intuitions vary between compatibilism and incompatibilism. In this section, we survey three of these factors.

(1) Explanation

Björnsson and Persson (2012a, 2012b) have developed what they call the Explanation Hypothesis, according to which an agent is morally responsible for an outcome only if "a *relevant motivational structure* of the agent is part of a *significant explanation* of the event" (2012a). Now, the problem is that whether something is part of a significant explanation depends on what counts as a significant explanation, and this is very likely to vary with context and background assumptions, and thus to be sensitive to framing effects.

More particularly, they argue that a range of incompatibilist arguments work only because they lead us to shift our focus from the agent's dispositions to a very distant past that caused these dispositions. In a deterministic universe, both the dispositions and their distant sources are part of an integral and complete explanation of the agent's behavior. Nevertheless, since it is hard for us to focus on both these distant sources and the actual dispositions, drawing attention on the distant sources elicits the feeling that the agent's actual dispositions do not explain his action, and thus that is not morally responsible for it. But this means that these incompatibilist arguments can elicit the incompatibilist intuitions only if we take the explanation by the distant sources as a more significant explanation than the explanation by the agent's dispositions. Thus, we can ask ourselves whether we *should* take the explanation by the distant sources as the most significant explanation.

[14] A recent study (Feltz, 2013) also suggests that Pereboom's manipulation argument does not work because people tend to judge a manipulated agent responsible as long as the manipulation is not intentional.

(2) Self

In a recent paper, Knobe and Nichols (2011) argue that people judge an act free only if it has been caused by the self and that people see determinism as a threat to free will only to the extent that they perceive it as implying that the self is not the source of our actions. They distinguish three conceptions of the self:

i. *The bodily conception of the self*, according to which the one's self is one's body.
ii. *The psychological conception of the self*, according to which one's self is one's mental states.
ii. *The executive conception of the self*, according to which the self is really some further thing, something over and above the various mental states one might have.

Knobe and Nichols' hypothesis is that people switch from one conception of the self to another depending on the context. This implies that people sometimes adopt the executive conception of the self — a conception according to which the self is not reducible to and determined by one's current mental states, but the source of its own actions. Such a conception is likely to trigger incompatibilist intuitions in situations that make it salient[15]. These different conceptions of the Self could then lead people to have incompatible intuitions[16].

(3) Individual differences

Finally, a limitation of NCH comes from the fact that, in all the studies we described, at least a minority of participants gave incompatibilist answers. Though one may be tempted to attribute these answers to noise or to participants failing to understand the vignettes, one possibility is that a minority of participants have genuine incompatibilist intuitions, making the claim that we are all natural compatibilists a case of overgeneralization.

This possibility has been explored by Adam Feltz and Edward Cokely, whose leading hypothesis is that philosophical intuitions might depend on stable and heritable character traits. Studying intuitions about free will and determinism, they

[15] For an argument that such intuitions in fact stem from a confusion, see Cossara (2012).

[16] In a related way, Sripada (2010) argues that free agents are not those who act according to their desires, but those who act according to desires expressing their 'Real' or 'Deep' self (their deepest commitments), though he adopts a compatibilist position according to which one can act according to one's 'Deep' self even if one lives in a deterministic universe. Susan Wolf (1987) had objected to this kind of views that it is not sufficient to act according to one's 'Real Self' to be free and that a person acting according to her 'Real Self' but who did not have the possibility to discover other values during her development was not really free. This position was based on appeal to intuition about particular cases. Faraci and Shoemaker (2010) have empirically tested Wolf's claims but found that people's intuitions did not support Wolf's objection to the 'Real Self' view.

found that people high in extraversion were more likely to judge an agent in a deterministic world as free and responsible for his actions (Feltz & Cokely, 2009).

One of Feltz and Cokely's conclusion is that "findings about individual differences might in part explain why the "free will problem" is such a persistent and intractable philosophical problem. It might be that different philosophers, because they have different personalities, motivations, and sensitivities, simply experience different intuitions about free will and moral responsibility's compatibility with determinism. It follows from research in the psychology of intuitive judgment that these differences in intuitions are likely to be mediated by different judgment processes. Thus, these different processes could generate different intuitions about the same philosophical example" (p.348). But, if this is true, then showing that most people have compatibilist intuitions is just bad news for incompatibilism without being good news for compatibilism, for it would be that intuitions cannot provide support for either position if they vary from one person to another.

4. Conclusion

In this paper, we sought to address the compatibility question by surveying both the experimental studies directly testing the intuitions in question and the ones testing the premises of the arguments for incompatibilism. There are still three remaining questions: how types of explanation, notions of the self and individual differences influence laypeople's intuitions. But, the balanced evidence from the experimental literature seems to us to be, so far, in favor of the view that *most* people have compatibilist intuitions; and that apparent incompatibilist answers can be explained away as the result of a misunderstanding of what determinism really is. However, these results are merely temporary, since other studies can overthrow current theories. Many incompatibilist arguments, like the Direct Argument, have not been submitted to empirical scrutiny, and showing that one relies on intuitive premises is enough to shed doubt on NCH. Thus, the jury is still out, though it might be leaning towards the 'natural compatibilists' conclusion.

Acknowledgements

We are grateful to Jesse Prinz (City University of New York) for his very helpful commentary on an earlier version of this paper. This research was supported by the National Center of Competence in Research (NCCR) Affective sciences financed by the Swiss National Science Foundation (no 51NF40-104897) and hosted by the University of Geneva.

References

Björnsson, G. & Persson, K. (2012a) The explanatory component of moral responsibility.

Noûs, 46, 326–354.

Björnsson, G. & Persson, K. (2012b) A unified empirical account of responsibility judgments. *Philosophy and Phenomenological Research*, 87, 611–639.

Cokely, E. & Feltz, A. (2009) Adaptative variation in judgment and philosophical intuition. *Consciousness and Cognition*, 18, 355–357.

Cossara, S. (2012) Cognitive science, moral responsibility and the self. *Baltic International Yearbook of Cognition, Logic and Communication*, 7, Article 3.

Cova, F. (2011) Neurosciences et droit pénal: le déterminisme peut-il sauver la conception utilitariste de la peine? *Klesis*, 21, 33–77.

Cova, F. (2012) Qu'est-ce que la philosophie expérimentale? In: Cova, F., Dutant, J., Machery, E., Knobe, J., Nichols, S. & Nahmias, E. eds. *La Philosophie Expérimentale*, Vuibert.

Cova, F., Bertoux, M., Bourgeois-Gironde, S. & Dubois, B. (2012) Judgments about moral responsibility and determinism in patients with behavioural variant of frontotemporal dementia: Still compatibilists. *Consciousness and Cognition*, 21, 851–864.

Cova, F. (forthcoming) Frankurt-style cases user manual: Why Frankfurt-style enabling cases do not necessitate tech support. *Ethical Theory and Moral Practice*.

De Brigard, F., Mandelbaum, E. & Ripley, D. (2009) Responsibility and the brain science. *Ethical Theory and Moral Practice*, 12, 511–524.

Faraci, F. & Shoemaker, D. (2010) Insanity, deep selves, and moral responsibility: the case of JoJo. *Review of Philosophy and Psychology*, 1, 319–332.

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

Feltz, A. & Cokely, E.T. (2009) Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition*, 18, 342–350.

Feltz, A., Cokely, E. & Nadelhoffer, T. (2009) Natural compatibilism versus natural incompatibilism: back to the drawing-board. *Mind and Language*, 24, 1–23.

Feltz, A. & Cova, F. (submitted) Moral responsibility and free will: A meta-analysis.

Frankfurt, H.G. (1969) Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66, 829–839.

Greene, J.D. & Cohen, J.D. (2004) For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London B*, 359, 1775–1785.

Harris, S. (2012) *Free Will*, S&S International.

Knobe, J. & Nichols, S. (2008) An experimental philosophy manifesto. In: Knobe, J. & Nichols, S. eds. *Experimental Philosophy*, Oxford University Press, 3–14.

Knobe, J. & Nichols, S. (2011) Free will and the bounds of the self. In: Kane, R. ed. *The Oxford Handbook of Free Will*, Oxford University Press.

Mandelbaum, E. & Ripley, D. (2012) Explaining the abstract/concrete paradoxes in moral psychology: the NBAR Hypothesis. *Review of Philosophy and Psychology*, 3, 351–368.

Miller, J. & Feltz, A. (2011) Frankfurt and the folk: An empirical investigation. *Consciousness and Cognition*, 20, 401–414.

Monroe, A.E. & Malle, B. F. (2010) From Uncaused Will to Conscious Choice: The Need to Study, Not Speculate About People's Folk Concept of Free Will. *Review of Philosophy*

and Psychology, 1, 211–224.

Murray, D. & Nahmias, E. (in press) Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research.*

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

Nahmias, E., Coates, J. & Kvaran, T. (2007) Free will, moral responsibility, and mechanism: Experiment on folk intuitions. *Midwest Studies in Philosophy*, 31, 214–242.

Nahmias, E., Morris, S., Nadelhoffer, T. & Turner, J. (2004) The phenomenology of free will. *Journal of Consciousness Studies*, 11, 162–179.

Nahmias, E., Morris, S., Nadelhoffer, T. & Turner, J. (2005) Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, 18, 561–584.

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

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

Nichols, S. (2006) Folk intuitions on free will. *Journal of Cognition and Culture*, 6, 57–86.

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

Pereboom, D. (1995) Determinism *al dente*. *Noûs*, 29, 21–45.

Roskies, A.L. & Nichols, S. (2008) Bringing moral responsibility down to earth. *Journal of Philosophy*, 105, 371–388.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S. & Sirker, (2010) Is belief in free will a cultural universal? *Mind and Language*, 25, 346–358.

Sinnott-Armstrong, W. (2008) Abstract + Concrete = Paradox. In: Knobe, J. & Nichols, S. eds. *Experimental Philosophy*, Oxford University Press, 209–230.

Sommers, T. (2010) Experimental philosophy and free will. *Philosophy Compass*, 5, 199–212.

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

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

Strawson, G. (1994) The impossibility of moral responsibility. *Philosophical Studies*, 75, 5–24.

Weigel, C. (2011) Distance, anger, freedom: an account of the role of abstraction in compatibilist and incompatibilist intuitions. *Philosophical Psychology*, 24, 803–823.

Wolf, S. (1987) Sanity and the metaphysics of responsibility. In: Schoeman, F. ed. *Responsibility, Character and the Emotions*, Cambridge University Press, 45–64

Woolfolk, R. (2011) Empirical tests of philosophical intuitions. *Consciousness and Cognition*, 20, 415–416.

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

(Received 2013.5.26; Revised 2013.12.19; Accepted 2014.1.20)