Why compatibilist intuitions are not mistaken: A reply to Feltz and Millan*

James Andow (Reading)
Florian Cova (Geneva)

April 13, 2015

Abstract

In the past decade, a number of empirical researchers have suggested that laypeople have compatibilist intuitions. In a recent paper, Feltz and Millan (in press) have challenged this conclusion by claiming that most laypeople are only compatibilists in appearance, and are rather willing to attribute free will *no matter what*. As evidence for this claim, they have shown that an important proportion of laypeople still attribute free will to agents in fatalistic universes. In this paper we first argue that Feltz and Millan's error-theory rests on a conceptual confusion: it is perfectly acceptable for a certain brand of compatibilists to judge free will and fatalism to be compatible, as long as fatalism does not prevent agents from being the source of their actions. We then present the results of two studies showing that laypeople's intuitions are best understood as following a certain brand of source compatibilism rather than a "free-will-no-matter-what" strategy.

Keywords. compatibilism; determinism; experimental philosophy; free will; moral responsibility

1 Introduction

Imagine that we discover all there is to know about human psychology, so that we understand everything about human behavior and its causes. Imagine that it turns out that human behavior is completely determined, in the sense that each human action and decision is completely caused by prior events that are not within human control, and that, given these prior events, it was necessary for this action or decision to happen. What would this mean for human free will and moral responsibility? *Compatibilists* think that determinism, understood in this sense, does not preclude free will and moral responsibility. *Incompatibilists*, however, consider that such a discovery would entail that there is no such thing as free will or moral responsibility, for free will and moral responsibility are incompatible with determinism. Compatibilists and incompatibilists disagree on what we will call the *compatibility question*.

How are we to decide between these options? As we indicated, no amount of empirical investigation of human behavior will be enough to decide between them. Compatibilists and incompatibilists' disagreement is not over what behavior human beings exhibit nor over the causes of human actions. They are not concerned with what *actually* causes human actions, but what human actions *should* be caused by to count as free. so, to answer the compatibility question we need to scrutinize the very nature of free will and moral responsibility.

But how do we do that? So far, philosophers have relied mainly on appeal to intuitions about free will and moral responsibility. On the incompatibilist side, arguments have relied first on the famous Principle of Alternate Possibilities, then on more complex intuitive principles, such as Van Inwagen's Rules B and β (Van Inwagen, 1983). Similarly, Arguments from Manipulation take intuitions as their starting point: in this case, intuitions about manipulation cases, coupled with the intuition that there is no relevant difference between such cases and cases involving agents in deterministic universes (see, e.g., Pereboom, 1995). Compatibilists have also made much progress using intuitions. They have used intuitions triggered by so-called Frankfurt cases (Frankfurt, 1969), and

*This is a prepublication copy of a paper forthcoming at *Philosophical Psychology*. Please cite the final version.

their responses to incompatibilist arguments have involved a whole industry of trying to create cases which pump intuitions that undermine these arguments' supposedly intuitive premises (e.g., Ravizza, 1994). Granted, not all attempts rely on intuitions, e.g., one research program approaches the question from the point of view of reactive attitudes (Strawson, 1962). However, it would be hard to deny that intuitions play an important role in inquiries about the compatibility question, whether because one thinks that what matters is our ordinary concept of free will, or because one thinks such intuitions give us access to some mind-independent essence of free will.

However, there is disagreement about what exactly our intuitions about free will and moral responsibility are, and how widespread they are. *Natural compatibilists* claim that laypeople pretheoretically tend to have compatibilist intuitions, while *natural incompatibilists* claim that laypeople enter the debate with incompatibilist intuitions (Feltz, 2009). In recent years, both sides have drawn on empirical investigations of laypeople's intuitions (for a review, see Cova and Kitano, forthcoming). However, these studies have yielded seemingly inconsistent results; people give compatibilist answers in some contexts, and incompatibilist answers in others. Both sides have thus developed error-theories aiming to explain why certain answers should be understood as the product of biases and confusions, rather than be taken seriously, or why apparent compatibilist (or incompatibilist) answers do not reflect genuine commitment to compatibilism (or incompatibilism).

In this paper, we investigate a recent proposal according to which seemingly compatibilist answers do not express genuine compatibilist intuitions, because they are the product of a tendency to judge agents to have free will 'no matter what'. After having situated this proposal within the current experimental literature on free will, we argue that this proposal rests on a failure to understand that certain contemporary brands of compatibilism allow for an agent to act freely in certain types of universe which might be branded fatalistic. We report results from two studies suggesting that 'compatibilist' answers, rather than following a dumb 'free will no matter what' pattern, are in fact sensitive to important differences between types of fatalist universe, and thus indicative of genuine compatibilist beliefs. Specifically, our results suggest participants are sensitive to whether a fatalistic universe involves the *bypassing* of agents' mental states—this notion will be explained below—rather than simply being a universe in which events were bound to happen.

2 Error theories in experimental philosophy of free will

Most studies of intuitions about free will and moral responsibility have focused directly on intuitions about whether an agent living in deterministic universe can be free and morally responsible. The first results were mostly in favor of natural compatibilism. In one of the first studies, Nahmias et al. (2006) asked participants to imagine that, in the next century, humans build a supercomputer able to accurately predict future human behavior on the basis of the current state of the world. Participants were then asked to imagine that, in this future, an agent has robbed a bank, as the supercomputer had predicted before he was even born. In this case, 76% of participants answered that this agent acted of his own free will, and 83% answered that he was morally blameworthy. These results suggest that most participants have compatibilist intuitions, since most answered that this agent could act freely and be morally responsible, despite living in a deterministic universe.

However, not all results have been so clear. In a subsequent study, Nichols and Knobe (2007) presented participants with the description of two different universes: Universe A, in which "everything that happens is completely caused by whatever happened before it", including human decisions, so that each decision "had to happen" the way they did, and Universe B, in which human decisions are not completely caused by whatever happened before, and did not have to happen the way they did.

After reading these descriptions, participants were assigned to one of the two experimental conditions. Participants in one condition – 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?

Participants in the other condition – the ABSTRACT CONDITION – however, 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 the CONCRETE CONDITION, most participants (72%) gave the compatibilist answer according to which the agent was fully morally responsible. In the ABSTRACT CONDITION , however, most participants (86%) gave the incompatibilist answer. We thus have apparently contradictory results. One solution is to accept the idea that people are neither pure compatibilists nor pure incompatibilists, but a bit of both (Weigel, 2013). Another is to propose an error-theory, i.e. a theory that explains how participants can be genuine compatibilists or incompatibilists while giving seemingly incoherent answers.

2.1 Nichols and Knobe's 'Performance Error Model'

The first error-theory advanced was Nichols and Knobe (2007)'s 'Performance Error Model', according to which people nat¬urally tend towards incompatibilism, but can be biased towards giving compatibilist answers by their affective responses to hateful crimes. According to this model, participants reveal their true commitment to incompatibilism in the ABSTRACT CONDITION, but are biased towards compatibilist answers by the outrage and indignation triggered by the crime described in the CONCRETE CONDITION. Though interesting, this proposal faces three major problems. First, Nahmias et al. (2006) have also found high rates of compatibilist answers in vignettes involving only neutral actions. Second, Cova et al. (2012) have found that patients suffering from a behavioural variant of frontotemporal dementia, a neurodegenerative disease accompanied by a deficit in emotional responses, were no less likely to give incompatibilist answers. Finally, a recent meta-analysis of 30 published and unpublished studies showed that, if affect had an effect on ascriptions of free will, this effect was far too small to explain the difference between the ABSTRACT and CONCRETE conditions (Feltz and Cova, 2014).

2.2 Murray and Nahmias' error-theory for incompatibilist intuitions

A second error-theory was proposed by Murray and Nahmias (forthcoming). Murray and Nahmias's key claim is that seemingly incompatibilist answers might not reveal a deep commitment to incompatibilism, but rather might be due to a conflation of determinism and what they call 'bypassing'. 'Bypassing' occurs when agents' mental states no longer play a role in the production of their actions. For example, if one acts under some hypnotic suggestion, one's mental states are 'bypassed'; what one really wants no longer causes one's action. It is important to distinguish bypassing from determinism. Both compatibilists and incompatibilists agree that free will and moral responsibility are impossible whenever bypassing obtains. However, only incompatibilists would claim that *determinism* makes free will and moral responsibility impossible. Thus, according to Murray and Nahmias, when participants give the incompatibilist answer in Nichols and Knobe's ABSTRACT CONDITION, it is not because they think that determinism and moral responsibility are incompatible, but rather because they think the universe described in Nichols and Knobe's vignette entails bypassing. In the CONCRETE CONDITION, however, this confusion between determinism and bypassing is lessened by the fact that the vignette makes it clear that the agent kills his family because he *wants* to be with his secretary, thus making clearer that the agent's mental states played a role in the production of his action.

To test for this hypothesis, Nahmias and Murray gave participants different scenarios describing agents acting in deterministic universes, including Nichols and Knobe's ABSTRACT and CONCRETE conditions. In addition to questions about morally responsibility and free will, participants received questions designed to probe to what extent they interpreted the vignette as implying some kind of bypassing (e.g. to what extent they agreed that "what the agent wanted had no effect on what he ended up being caused to do").

Nahmias and Murray's results matched their predictions. First, they found that compatibilist intuitions were highly correlated with a good understanding of determinism (that is: one which doesn't conflate it with bypassing) and that participants who gave incompatibilist answers were far more susceptible to the belief that determinism entailed bypassing. Second, they found that people were more likely to read the vignette as implying some kind of bypassing in the ABSTRACT than in the CONCRETE condition. They conclude that most incompatibilist answers are only apparent incompatibilist answers, because most participants who give incompatibilist answers chose them to express the intuition that free will and moral responsibility are incompatible with bypassing—something with which compatibilists would agree—rather than determinism.

2.3 Feltz and Millan's error-theory for compatibilist intuitions

So, we have reason to believe that most incompatibilist answers to such vignettes are the product of a misled interpretation of such vignettes as implying bypassing. However, this does not mean that compatibilist answers involve no similar confusion. Feltz and Millan (forthcoming) argue that most 'compatibilist' answers are not best interpreted as indicating that participants are genuinely compatibilist, because people who give them seem to have 'free-will-no-matter-what' (FWNMW) intuitions. In other words, most people who give 'compatibilist' answers, in response to the vignettes we have described, still judge agents to be free and morally responsible for their actions when presented with other vignettes in which it is clear for most philosophers, including compatibilists, that the agents described cannot act freely nor be morally responsible for their actions. Thus, Feltz and Millan conclude, it is unlikely such participants have proper compatibilist intuitions.

Before moving on, note that the error-theories we have considered so far are not error-theories in the same sense. Nichols and Knobe's account seems to consider that people genuinely believe an agent can be morally responsible in a deterministic universe, and thus give genuine compatibilist answers, when faced with high-affect cases. However, these answers are contrary to what participants would think and answer if they reflected in a cool and rational way. In this case, so the error-theory goes, participants are genuinely biased: they give genuine answers that contradict their inner standards. Nahmias and Murray's error-theory is rather different; it says that people who give seemingly incompatibilist answers because they think a case implies bypassing, are giving answers that are coherent and derivable from their inner commitment to compatibilism. This involves no error on the participants' part. If it did, it would have to be mistakenly judging the vignette to describe an agent whose mental states are bypassed. However, nothing in the vignette precludes such an interpretation so this can't really be said to be an error made by participants. Rather, according to Nahmias and Murray's error-theory, it is the experimenters who are mistaken; their mistake, interpreting answers motivated by compatibilist intuitions as manifesting incompatibilist intuitions. So, our first two error-theories are error-theories in distinct ways.

Feltz and Millan's error-theory is of the second kind. Besides compatibilist and incompatibilist intuitions, they distinguish a third kind of intuitions: 'free-will-no-matter-what' intuitions. Then, they argue that seemingly compatibilist answers are in fact FWNMW intuitions that have been mistakenly interpreted as compatibilist.

However, one might wonder why Feltz and Millan distinguish FWNMW intuitions from 'genuine' compatibilist intuitions. After all, people with FWNMW intuitions think that agents living in deterministic universes can act freely and be morally responsible for their actions, and this is what we expect a compatibilist to answer. So, why deny that FWNMW intuitions are 'genuine' compatibilist intuitions? The answer is that compatibilist philosophers are not only in the business of arguing that free will (or moral responsibility) and determinism are compatible; they also aim at developing a plausible account of free will and moral responsibility. As such, compatibilist philosophers do not think that agents can be free and morally responsible 'no matter what'. Rather, they argue (i) that agents must exert a certain kind of control on their action to be free and morally responsible, and (ii) that it is possible to exert this kind of control even if determinism is true, and thus conclude (iii) that free will and moral responsibility are compatible with determinism. People with FWNMW intuitions share with compatibilist philosophers their conclusion (iii) but do not seem concerned about the issue of control. Rather, they seem to follow a crude heuristic according to which agents are morally responsible whatever the circumstances. This is where we can draw the line between genuine compatibilist intuitions and FWNMW intuitions: while genuine compatibilist intuitions are sensitive to the control agents have over their actions, FWNMW intuitions are not. As Feltz and Millan put it, "to have genuine compatibilist intuitions, people need to be able to distinguish between cases where an individual satisfies compatibilist necessary conditions for freedom and moral responsibility (e.g., some conditions which require agents actions and mental states to be causally integrated in an appropriate way), while being able to recognize instances where a person does not satisfy those conditions".[1]

Now, Feltz and Millan argue that most people who give seemingly compatibilist answers to the kind of vignette we described above are not genuine compatibilists but rather have FWNMW intuitions. Is this really the case? In the next section, we detail and discuss Feltz and Millan's argument in favor of this conclusion.

[1]The notion of 'free will no matter what' intuitions originates in Feltz (2009). There the focus is on participants who have an entrenched commitment to a worldview which makes it difficult for them to imagine/understand a description of a deterministic universe (Feltz, 2009, 16–17). In this paper, however, we follow Feltz and Millan in classifying as having 'free will no matter what' intuitions those participants who 'would judge that an agent is free and morally responsible even in fatalistic scenarios' (p.6) or in any other case in which it would be obvious even to the most die-hard compatibilist that one cannot be free and morally responsible.

3 Fatalism, Frankfurt cases, and compatibilism: what factors should compatibilist intuitions be sensitive to?

So, how could one show that apparent compatibilist answers are in fact the expression of FWNMW intuitions? Feltz and Millan's strategy is to show that people who give apparently compatibilist answers in the case of an agent living in a deterministic universe also tend to judge this agent as free and morally responsible in a setting that should preclude free will and moral responsibility according to compatibilist philosophers.

3.1 Determinism vs. Fatalism

Adopting such a strategy requires picking out a context, different from determinism, in which free will and moral responsibility are impossible according to compatibilist philosophers. Accordingly, Feltz and Millan contrast *determinism* with *fatalism*, fatalistic universes supposedly being contexts in which free will and moral responsibility are impossible according to compatibilist philosophers.

Feltz and Millan define fatalism as follows: a fatalistic universe is a universe in which everything that happens *had to happen, regardless of what happened prior*.[2] Fatalistic universes can be contrasted with deterministic universes, which are universes in which everything that happens *had to happen, given what happened prior and the law governing the universe*. Let's take an example: James buys medicine at t2. In both deterministic and fatalistic universes, James *had* to buy medicine at t2. However, in a deterministic universe, James had to buy medicine at t2 *given the fact that* he fell ill at t1 (together with the laws governing this universe). Had something else happened at t1, and thus the initial conditions been different, then it would no longer have been true that James had to buy medicine at t2. Necessity in deterministic universes is relative to initial conditions and laws of nature: had any of these been different, then what happened no longer had to happen. However, in a fatalistic universe, James *had* to buy medicine at t2, regardless of what happened at t1. Even if James did not fall ill at t1, it would still be the case that James had to buy medicine at t2. Necessity in fatalistic universes is not relative but absolute.

Thus, according to Feltz and Millan, fatalistic and deterministic universes differ significantly: fatalism implies that we have no influence on what happens to us whereas determinism does not have that implication. Still, according to them, this means that though compatibilist philosophers claim that free will and moral responsibility can exist in a deterministic universe, they reject the possibility of free will and moral responsibility in fatalistic universes because fatalism is incompatible with having the sort of causal influence over actions that compatibilists require. Accordingly, Feltz and Millan probe whether participants have genuinely compatibilist intuitions by investigating whether participants who judge agents in deterministic universes to be free and morally responsible can see that agents in universes in which naïve fatalism holds are neither free nor morally responsible because such agents don't have the sort of causal influence over their actions that compatibilists require.

Feltz and Millan investigate this possibility through five different studies in which they compare participants' answers to vignettes describing deterministic and fatalistic universes. For example, in their fifth study, the compare the following deterministic scenario:

> Fries
> 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. 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.

[2] One might raise a number of issues about this way of defining fatalism. Because we follow Feltz and Millan in their use of 'fatalism', we take time address some of these issues and defend our usage in section 3.2.

To the following fatalistic scenario:

> Extreme Book
> There is a special book that has all of our decisions and actions truly written in its content. For instance, whenever we are trying to decide what to do, the decision we end up making is completely and truly written in this book and the decision will happen regardless of our thoughts, beliefs, desires, and plans. The special book has these events truly written in it lifetimes before the events took place. So, if the book has an event written in it, the event will definitely occur regardless of the past events and the laws of nature. For example, one day a person named John decides to kill his wife so that he can marry his lover, and he does it. Once the specific event is truly written in the book, it is impossible for John not to kill his wife regardless of his thoughts, beliefs, desires, and plans. Assume the book's contents made it impossible for John not to kill his wife. Please rate to what degree you agree with the following statements.

Participants received both cases. Feltz and Millan found that (i) participants were less likely to attribute free will and moral responsibility in fatalistic cases than in deterministic cases, but (ii) a substantial number of participants still considered the agent to be free and morally responsible in the fatalistic cases, and (iii) participants' answers in the fatalistic cases were a significant predictor of their answer in the deterministic cases. According to Feltz and Millan, this suggests that a substantial number of seemingly compatibilist participants in fact have FWNMW intuitions, for they attribute free will and moral responsibility both in the deterministic and fatalistic cases.[3]

3.2 On defining fatalism

It is important at this point to make a couple of notes on our use of the word 'fatalism'.[4] In the rest of this paper, we will be following Feltz and Millan's definition: a fatalistic universe is a universe in which everything that happens had to happen, regardless of what happened prior. This way of using the word 'fatalism' is somewhat unfortunate. The majority of philosophers in this area do not use the word in this way. Moreover, fatalism so-defined is not a particularly interesting philosophical position in itself. Let's first say a little more about more typical ways of using 'fatalism' and how it differs from our usage. We'll then take some time to explain why we use the term as we do.

'Fatalism' is typically used in the philosophical literature to refer to a more sophisticated position. For example, Van Inwagen (1983, 23) defines fatalism as "the thesis that that it is a logical or conceptual truth that no one is able to act otherwise than he in fact does; that the very idea of an agent to whom alternative courses of action are open is self-contradictory." Such a definition of fatalism does not imply that things will happen "no matter what we do". Rather, these more sophisticated brands of fatalism will hold that certain things had to happen, and that we also had to act a certain way. (All this said, there are some in the literature who seem to define fatalism in a similar way to Feltz and Millan, e.g., see Kane 2005, 19: "Fatalism is the view that whatever is going to happen, is going to happen, no matter what we do." One reason, for that might be that this use of "fatalism", though at odds with the contemporary philosophical use of "fatalism", seems to stem directly from its more traditional uses in modern philosophy and literature, e.g., see Diderot 1986.)

It thus might be useful to distinguish between *naïve fatalism* (certain things will happen no matter what we do) and *sophisticated fatalism* (things have to happen the way they do, but we also have to act the way we do). Feltz and Millan use the word "fatalism" to refer to naïve fatalism. Naïve fatalism holds that, if it was my fate to be cured from a particular disease, I will be cured no matter what I do – whether I consult a doctor or simply run naked in the snow. Sophisticated fatalism, on the contrary, holds that it was my fate to go to the doctor, be prescribed the relevant medicine, and be cured. Thus, naïve fatalism falls prey to the famous *argos logos* ('Lazy Argument') according to which one can do whatever one wants and this will make no difference, since what happens to us has

[3] One possible issue, which we won't deal with here, is the fact that both Extreme Book and Fries are likely not interpreted by participants as concerning the actual world, but rather present alternative possible worlds. This might have an effect on the results of studies using these scenarios as there is some evidence that participants' judgements about whether protagonists have free will are sensitive to whether the protagonist's universe is presented as being the actual universe (Roskies and Nichols, 2008). It would be interesting to probe this further in the context of 'free will no matter what intuitions' however we suspect this would prove difficult as participants would likely have difficulty imagining that the actual world is a fatalistic one—let alone one involving magic books.

[4] Thanks to an anonymous referee for pushing us on this point.

to happen whatever we do, while sophisticated fatalism does not. Finally, naïve and sophisticated fatalism do not differ from determinism in the same way: naïve fatalism differs from determinism to the extent that it deprives our action of causal significance, while determinism doesn't, and sophisticated fatalism differs from determinism to the extent that it claims that it is a logical or conceptual truth that one could not have done otherwise, while determinism does not.

Due to the fact that we follow Feltz and Millan in employing a somewhat atypical definition of fatalism, it might seem that we are open to the following objection: the empirical results we present cannot speak to the relevant philosophical debates because we adopt this definition. While such criticism is misguided, it will hopefully clarify our project to take some time to address it. The thought behind such a criticism might be this: if one wants one's empirical findings to apply to traditional philosophical debates, one has to operationalize the relevant key terms in concordance with the way philosophers have used these terms; if one doesn't do this, the data one obtains about folk intuitions won't speak to the relevant debates. This is a good thought which we endorse, and it might be a good criticism of an empirical study which aimed to cast light on a traditional philosophical debate about *fatalism* which adopted such an atypical understanding of fatalism. However, neither the current study (nor Feltz and Millan's) aims to cast light on debate about *fatalism*. Rather, the current study (and Feltz and Millan's) aims to cast light on the debate between compatibilists and incompatibilists. Let us reiterate what role intuitions about universes in which naïve fatalism holds play in Feltz and Millan's argument. Feltz and Millan are *not* aiming to find out to what extent laypeople have intuitions which accord with some philosophically interesting form of fatalism. Feltz and Millan's argument is as follows: (i) a certain type of universe, call them X-universes, is incompatible with free will and moral responsibility even under a compatibilist conception of free will; but (ii) an important number of participants judge agents in X-universes to be free and morally responsible; and thus (iii) an important number of participants are not genuine compatibilists but have FWNMW intuitions. Feltz and Millan call the relevant class of universes 'fatalistic universes' but the name is unimportant when it comes to the light their data and argument promise to shed on a traditional philosophical debate. Of course, as we noted previously, Feltz and Millan's use of "fatalism" is unfortunate. However, clearly their argument has the same value whatever term is used to describe the relevant class of universe (one might use, e.g. "naïve fatalism" or "bypassing universes"). Thus, the fact that Feltz and Millan do not use "fatalism" in the sense common in contemporary philosophy does nothing to undermine their argument, nor the way their studies are operationalized. Accordingly, the fact we follow Feltz and Millan's usage does nothing to undermine our study. The studies are designed to investigate participants' intuitions about what Feltz and Millan call "fatalism" and not their intuitions about what contemporary philosophers call "fatalism". (Of course, as a referee pointed out, this also means that they teach us nothing concerning what most philosophers call "fatalism", but this is not their purpose.)

3.3 Is compatibilism incompatible with fatalism?

Feltz and Millan's results suggest that many people who give seemingly compatibilist answers in deterministic cases also attribute free will and moral responsibility in fatalistic cases. Does this really mean that their intuitions are not genuine compatibilist intuitions? For Feltz and Millan, it does, because they contend that compatibilism is incompatible with the claim that free will and moral responsibility are possible in fatalistic universes. But why should we think that compatibilists are tied to the claim that free will and moral responsibility are impossible in fatalistic cases?

Feltz and Millan provide two reasons. The first is that fatalism is incompatible with the *ability to do otherwise*, even under a compatibilist understanding of this ability. Indeed, there has been a lot of debate about whether an agent in a deterministic universe has the ability to do otherwise is relevant for moral responsibility. Incompatibilists typically think that such an ability is incompatible with determinism, for it should be understood as the ability to do otherwise, *all prior conditions being kept the same*. Certain compatibilists, on the contrary, have argued that the relevant ability is compatible with determinism, for it should be understood as the ability to do otherwise provided that certain prior conditions (and, in particular, certain of the agent's mental states) *are different*. However, even under such a compatibilist interpretation, the ability to do otherwise is incompatible with fatalism, for fatalism implies that things have to happen the way they happen *whatever the prior conditions*.

The second reason Feltz and Millan provide is that, in fatalistic universes, it is impossible for agents not to do what they actually do, regardless of their mental states. To put it otherwise: in fatalistic contexts, agents' mental

states do not seem to matter, for things happen whatever mental states agents have. Since the relation between one's mental states and one's actions is at the core of most compatibilist accounts of free will and moral responsibility, and since fatalism seems to compromise this connection, then fatalism seems to undermine compatibilist conditions for free will and moral responsibility.

Thus, Feltz and Millan provide two reasons to believe that compatibilists would deny that agents in fatalistic contexts can be free and morally responsible: such contexts compromise one's ability to do otherwise, and they make agents' mental states impotent. Now, it is easy to imagine a case that shares both properties (Cova, 2014; Miller and Feltz, 2011):

> Car
> Mr. Green wants Mr. Jones, the security guard, to steal Mrs. Green's car at 12:00am on October 7. However, Mr. Green doesn't entirely trust Mr. Jones to do the job, so he has taken some extraordinary measures. Mr. Green has consulted neuroscientists who have implanted a device in Mr. Jones's brain without Mr. Jones's knowledge. This device has isolated the "decision-making" neurons in Mr. Jones's brain and is programmed to send, at exactly 12:00am, impulses that will certainly cause Mr. Jones to decide to steal the car just then. However, as it happens, at exactly 12:00am, Mr. Jones decides on his own to steal the car and does it. Since Mr. Jones decides on his own to steal the car, the impulses from the device were ineffectual because the decision-making neurons were activated by the decision-making process of Mr. Jones himself. However, if Mr. Jones had not, just then, decided on his own to steal the car, the device would have activated his decision-making neurons, and Mr. Jones would have decided to steal the car anyway.

In this case, Mr. Jones cannot do other than steal Mrs. Green's car *and* his mental states do not matter, in the sense that whatever his mental states, he would have ended up stealing Mrs. Green's car. Thus, this case can be considered as belonging to a particular species of fatalistic cases, in which there is only one action that *had* to happen: Mr. Jones' stealing of Mrs. Green's car. However, when presented with this case, most people answer that Mr. Jones acted of his own free will and is morally responsible for stealing Mrs. Green's car, while acknowledging that Mr. Jones could not have done otherwise.

Those who are familiar with the philosophical literature about free will and moral responsibility will probably have immediately recognized Car as being a Frankfurt-style case (Cova and Kitano, forthcoming; Frankfurt, 1969). Frankfurt-style cases are a very famous class of thought experiments designed to pump the intuition that one can act freely and be responsible for one's actions even if one could not have done otherwise. Of course, most incompatibilist philosophers deny that Frankfurt cases have actually showed that one can be free without having the ability to do otherwise, and even certain compatibilist philosophers (so-called 'leeway compatibilists') still think that the ability to do otherwise is a necessary condition for free will and moral responsibility. Still, Frankfurt-style cases have been widely influential and many compatibilist philosophers now accept the idea that one can be free and morally responsible if one could not have done otherwise.[5]

There have been various attempts to explain how an agent can be free and morally responsible even without the ability to do otherwise. One famous explanation draws on Fischer's distinction between two kinds of control: actual (or guidance) control and regulative control (Fischer, 1986). The distinction between the two kinds of control can be illustrated by the following example:[6]

> Al is taking a driver's education class. He is in the driver's seat operating one set of controls of the car. His teacher is Bart, who is carefully monitoring Al's driving, tells him to turn left. Al signals to make a left turn and proceeds to turn the steering wheel to the left, thus causing the car to make a left turn.

In this case, Al can be said to control the car's movement to the left. He has a certain sort of control of the car's

[5] Here we are not concerned with Frankfurt's actual view concerning free will and responsibility, but rather the effect that so-called 'Frankfurt cases' play in the literature.

[6] We take this example from Fischer and Ravizza (1991) as we think it nicely illustrates the distinction. Though Fischer and Ravizza ultimately settle for a form of semi-compatibilism, for which only moral responsibility – but not free will – is compatible with the lack of alternative possibilities, it is still possible to refuse to dissociate free will and moral responsibility, and use their argument in favor of a more classical form of compatibilism. For a review of data suggesting that people do not dissociate free will and moral responsibility, see Cova and Kitano (forthcoming).

turning to the left: *actual causal control*. This means that, insofar as Al deliberates in the normal way and there is no malfunction in the car's steering apparatus, Al can be said to have caused the car's turning left.

> Suppose that Al's teacher Bart has a dual set of controls of the car. If Bart wishes, he can activate his controls and deactivate Al's; thus, Bart can take control of the car, if he wishes. As things actually work out, Al controls the car's movement to the left and Bart plays no causal role in it. But, by virtue of the second set of controls, Bart has a dual power with regard to the car's turning to the left: he can ensure this event and he can prevent it. That is, if Al showed signs of wanting to turn the car to the right, Bart could override Al's attempt and cause the car to go to the left. Further, Bart could frustrate Al's attempt to cause the car to go to the left, if he wished; Bart could activate his controls and turn the car to the right.

In this case, Bart can both ensure that the car turns left and prevent it from turning left: he has *regulative control* over the car turning left. Because regulative control requires that one had the ability to prevent something that happened from happening or to make something that did not happened happen, it requires the ability to do otherwise. Imagine that Al turns the car left: it is still the case that Bart had regulative control over this event, since he could have prevented it from happening – he could have made things happen *differently*.

However, actual causal control (or guidance control) does not require the ability to do otherwise. Imagine that Al decides to turn left and, steering the wheel, causes the car to do so. But also suppose that, had Al decided not to go left, Bart would have taken control of the car and made it go left. It is true, in this case, that Al had no regulative control on the car turning left: had he chosen not to go left, Bart would have still made the car go left. Still, in the actual sequence, since it is Al who turns the car left, Al still has actual causal control on the car: he is the one who causes it to go left, even if he could not have prevented the car from going left.

Fischer's idea is that, though it is true that one has to have control upon a given event to be responsible for this event, one does not have to have regulative control: actual causal control is sufficient for moral responsibility. This is why agents in Frankfurt-style cases can be morally responsible; despite being deprived of regulative control, by the presence of a counterfactual intervener, they still exert actual control and they do so deliberately, on the basis of reasons.

Let's now apply Fischer's distinction to fatalistic settings, and more precisely to the Extreme Book case used by Feltz and Millan. It is true that, in Extreme Book, John does not have regulative control over killing his wife: given the very nature of the fatalistic setting, he had to, no matter what. However, this does not preclude John from having actual causal control, as long he causes his wife's death in the appropriate way (that is: not by coercion or by accident, but by acting on the basis of reasons). In fact, the text of Extreme Book suggests John has actual causal control. It is explicitly stated that John killed his wife "so that he can marry his lover"! Thus, it is not clear that John lacks free will and moral responsibility: a compatibilist account, according to which actual causal control is enough, would judge John to be morally responsible for killing his wife.

In fact, all compatibilist philosophers who accept that agents can be free and morally responsible in Frankfurt-style cases should accept that agents can be free and morally responsible in fatalistic universes, as long as agents are able to cause and produce their actions in normal ways (i.e., as long as bypassing does not obtain). Thus, just because participants judge agents in fatalistic universes to be free and morally responsible, does not mean that they do not have genuine compatibilist intuitions. It only means that they are not 'leeway compatibilists', i.e., the kind of compatibilists who think that the ability to do otherwise is necessary for free will and moral responsibility.

This suggests an alternate interpretation of Feltz and Millan's data: maybe the participants they claim to have FWNMW intuitions are what we shall call 'source compatibilists'. Source compatibilists do not think the ability to do otherwise is necessary for free will and moral responsibility, only that an agent is the source of her own actions. Source compatibilists should be willing to judge agents as possibly free and morally responsible as long as bypassing does not obtain. Maybe these source compatibilist participants tend to read the fatalistic cases as not implying bypassing, and thus appropriately judge the agent to be free and morally responsible. And maybe participants who do not read fatalistic cases as implying bypassing are also more likely to read deterministic cases as not implying bypassing either, which would explain the correlation Feltz and Millan observed between participants' answers to both cases. In the following sections, we put this alternate reading of Feltz and Millan's results to test.

4 Study 1: guidance control in a fatalist universe

4.1 Participants

129 participants living in the United States were recruited using Amazon's Mechanical Turk; those who completed the survey were paid $0.40.[7] 3 were excluded for having incomplete answers.[8] 37 were excluded for answering a comprehension question incorrectly.[9] Of the remaining 89 participants, 55 were men. Age mean was 33.56 (SD=12.86).

4.2 Materials

Participants received two scenarios in a randomized order. The scenarios were Feltz and Millan's Extreme Book and Fries.[10] Extreme Book describes a fatalistic universe, its protagonist was 'John'. Fries describes a deterministic universe, its protagonist was 'Bill'. After reading each text, participants had to rate their agreement with the following statements (on a scale from 1=disagree to 7=agree):

1. John's/Bill's killing of his wife was up to him.
2. John/Bill killed his wife of his own free will.
3. John/Bill is morally responsible for killing his wife.
4. John's/Bill's decision to kill his wife and children had no effect on what he ended up being caused to do.
5. What John/Bill wanted had no effect on what he ended up being caused to do.
6. What John/Bill believed had no effect on what he ended up being caused to do.
7. John/Bill had no control over what he did.

These seven items incorporate two distinct pre-established scales. Statements 1–3 comprise a scale which aims to assess the degree to which participants judged the agent to have free will and moral responsibility. These are standard items in the experimental literature (see, e.g., Cova et al., 2012). Statements 4–7 comprise a scale which aims to assess the degree to which participants interpret the relevant scenario as entailing bypassing. These items were used by Murray and Nahmias (forthcoming) as a scale to calculate a composite bypassing score (as we shall below).[11] After answering these questions, participants were asked to answer the corresponding comprehension question. After having answered questions for both scenarios, participants completed the Ten Item Personality Inventory (Gosling et al., 2003).[12] Finally, basic demographic information was gathered.

4.3 Results

First, we checked whether our results were consistent with Feltz and Millan's data. Means and standard deviations for the 7 statements are reported in Table 1. There was a strong internal consistency among responses to statements 1-3 for both Fries (Cronbach's α=0.95) and Extreme Book (α=0.89). Internal consistency was also strong

[7] This doesn't include 14 participants who left the survey without answering a single question.

[8] These participants responded to only one of the two scenarios.

[9] As Feltz and Millan note (n.8, p. 24), high rates of comprehension failures are the norm in experiments on free will; we excluded around 29% of participants who gave complete answers. Feltz and Millan draw attention to Sommers (2010) who "notes that people routinely exclude 10–30% of participants in these types of studies". These exclusions are necessary, however, to ensure participants understand Extreme Book and Fries adequately. The standard comprehension questions used were (i) 'If the universe were re-created with the special book having the same true sentences, John would do the same thing?' (Extreme Book) and (ii) 'If Universe A was exactly recreated, is it accurate to say that Bill would do the same thing?' (Fries). Participants were encouraged to take their time in the survey. They were told their answers would not be approved (in MTurk) if they spent less than three minutes on the survey.

[10] Though we used the same texts as Feltz and Millan, two minor changes were made: (i) there was someone called John in both scenarios and, as this could be confusing in a within subjects design, in Fries John became Frank; (ii) the opening of our Extreme Book asked participants to 'Imagine that there is a special book…'.

[11] For a discussion of this scale's validity, see Björnsson and Pereboom (2014). One issue is that item 7 is potentially ambiguous. It *could* be read as asking about whether the protagonist has actual control or whether they have regulative control (see section 3.2). This obviously bears on the extent to which the scale measures attitudes to bypassing and we are sympathetic to Feltz and Millan's suggestion that use of methods such as protocol analysis may be needed to understand the precise nature of participants' intuitions in this area.

[12] We were not personally interested in individual differences. However, Feltz and Millan gave the TIPI to their participants and we followed them to stay as close to their original experimental design as possible.

among responses to statements 4-7 for both Fries (α=0.86) and Extreme Book (α=0.9). A composite score was calculated for free will and moral responsibility (FWMR), averaging answers to statements 1-3, and for bypassing (BYPASS), averaging answers to statements 4-7.[13]

Table 1: Mean answers (and standard deviations) to statements 1-7 for Fries and Extreme Book cases in Study 1.

		Extreme Book		Fries	
	Statement	M	SD	M	SD
1	John's/Bill's killing of his wife was up to him	2.99	2.18	4.48	2.40
2	John/Bill killed his wife of his own free will	3.01	2.22	4.42	2.38
3	John/Bill is morally responsible for killing his wife	3.80	2.34	4.97	2.32
4	John's/Bill's decision to kill his wife and children had no effect on what he ended up being caused to do	4.30	2.15	3.45	2.26
5	What John/Bill wanted had no effect on what he ended up being caused to do	4.64	2.27	3.42	2.27
6	What John/Bill believed had no effect on what he ended up being caused to do	4.72	2.12	3.58	2.25
7	John/Bill had no control over what he did.	4.80	2.20	3.53	2.43

FWMR ratings were higher in the deterministic case (Fries) than the fatalistic case (Extreme Book). A mixed model ANOVA was conducted with the two FWMR scores as within-subjects factors and order of presentation as a between-subjects factor. This indicated subjects rated Fries (M=4.62, SD=2.25) higher than Extreme Book (M =3.27, SD=2.04) on FWMR ratings (F(1,87)=26.99, p<.0005, η_p^2 = .24). Order did not interact with judgments (p=.586). Then, multiple linear regression was conducted with order of presentation and Extreme Book FWMR ratings as predictor variables and Fries FWMR ratings as the dependent variable. The full model was a significant predictor of Fries FWMR ratings (F(2,86)=7.08, p=.001). Thus, participants' FWMR ratings in the fatalistic case predicted their FWMR ratings in the deterministic case. After controlling for order of presentation, Extreme Book FWMR ratings continued to predict Fries FWMR ratings (β =.362, t=3.622, p<.0005). Addition of Extreme Book FWMR ratings to a model containing only order of presentation was associated with a R^2_{change} of .131.

Thus, we reproduced Feltz & Millan's results. We turned our attention to the relation between BYPASS and FWMR scores. First, a mixed model ANOVA was conducted with the two BYPASS scores as within-subjects factors and order of presentation as a between-subjects factor was conducted. Participants BYPASS ratings were lower for Fries (M=3.49, SD=1.92) than Extreme Book (M=4.62, SD=1.91) (F(1,87)=21.19, p<.0005, η_p^2=.2; order did not interact, p=.71). However, there was a correlation between participants' BYPASS ratings in both cases (r(87) =.281, p=.008). Thus, participants who had higher BYPASS scores in the Extreme Book case were also more likely to have high BYPASS scores in the Fries case. So, BYPASS scores behave exactly like participants' FWMR scores.

Could it be, then, that BYPASS scores predict FWMR scores and that participants' judgments about free will and moral responsibility are explained by their reading of the different cases as implying or not implying bypassing? This is further suggested by a strong correlation between FWMR and BYPASS scores in both Extreme Book (r(87)=-.672, p<.0005) and Fries (r(87)=-.645, p<.0005). To further study this relationship, we conducted a midpoint split on FWMR and BYPASS scores, as illustrated in Table 2. This suggests a clear pattern: the difference in FWMR ratings between Fries and Extreme Book seems to be accounted for by participants' answers to the BYPASS questions. We treated composite scores above the midpoint of each scale (i.e., 4) as agree (either that bypassing obtains, or that protagonists have free will) and those below as disagree (any with composite scores of exactly 4 are not included; percentages are of the total number of participants).

[13]One item (3 in Extreme Book) showed a significant correlation with age r(87)=.3, p=.004. There were no gender effects. We looked at the relation between personality and FWMR scores. We used the personality traits Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experiences (ascertained using the Ten Item Personality Inventory) and BYPASS scores to predict FWMR scores in linear regression. The full model was a significant predictor in both Fries (F(6,82)=13.052, p<.0005, R^2=.488) and Extreme Book (F(6,82)=14.264, p<.0005, R^2=.511). Controlling for BYPASS and the other personality traits, Extraversion continued to predict FWMR in Fries (β=.217, t=2.395, p=.019) (it did not for Extreme Book). This fits with Feltz and Millan' findings, and with the literature more widely.

Table 2: Midpoint split of FWMR & BYPASS scores for each scenario in Study 1.

			BYPASS	
			Disagree	Agree
FWMR	Extreme Book	Disagree	13.5%	50.6%
		Agree	22.5%	9%
	Fries	Disagree	11.2%	23.6%
		Agree	47.2%	11.2%

Table 3: Multiple linear regression analysis with order of presentation, scenario-type and BYPASS scores as predictors and FWMR scores as the dependent variable for Study 1.

Model	B	SE	β	P
Constant	7.185	.296		<.0005
Order	.05	.247	.011	.839
BYPASS	-.74	.065	-.654	<.0005
Fatalist/Determinist	-.524	.255	-.117	.04

Indeed, BYPASS scores were generally good predictors of FWMR scores. We conducted multiple linear regression with order of presentation, scenario-type (coded as follows: extreme book=1, bypassing=0) and BYPASS scores as predictor variables and FWMR scores as the dependent variable.[14] The full model was a significant predictor of FWMR score (F(3,174)=54.34, p<.0005) and accounted for approximately 50% of the variance (R^2=.484, adj. R^2=.475). Significantly, BYPASS scores were a more important predictor than whether the scenario was fatalist or determinist (see Table 3 for full model). When scenario-type was added to the model in a final block, it was associated with only a minor improvement (R^2_{change}=.013).

Finally, although our results were consistent with Feltz and Millan's data to the extent that Fries FWMR scores were predicted by Extreme Book FWMR scores, we found that participants' Fries BYPASS responses were more important predictors. This is important. If participants really attribute free will no matter what, it should be the case that participants' compatibilist intuitions (and FWMR ratings) in Fries are best predicted by their FWMR score in Extreme Book (rather than BYPASS score in Fries). Whereas, if participants are source compatibilists and so sensitive to whether agents exert actual causal control, it should be the case that participants' compatibilist intuitions (and FWMR ratings) in Fries are best predicted by their BYPASS score in Fries. Multiple linear regression with order of presentation, Extreme Book FWMR scores and Fries BYPASS scores as predictor variables and Fries FWMR scores as the dependent variable was conducted. The full model was a significant predictor of Fries FWMR (F(3,85) = 25.85, p<.0005). The full model is presented in Table 4. It is true that, after controlling for order of presentation, Extreme Book FWMR scores continued to predict compatibilist judgments about Fries (β=.25, t=3.125, p=.002). However, addition of Extreme Book FWMR scores to a model containing Order and Fries BYPASS scores was associated with a small R^2_{change} of .06. On the other hand, the Fries BYPASS scores was a more important predictor (β=-.593, t=-7.387, p<.0005), and addition of Fries BYPASS scores to a model containing Order and Extreme Book FWMR scores was associated with a R^2_{change} of .336.

4.4 Discussion

Feltz and Millan argued on the basis of their results that many participants who consider the protagonist in Fries to have free will should not be interpreted as genuine compatibilists. According to them, these participants have FWNMW intuitions instead: they are also willing to say that the protagonist in Extreme Book has free will, which is supposed to show that their intuitions do not reflect a genuine folk compatibilism. As we noted, these participants could well be *source compatibilists* so long as they do not read Extreme Book as entailing bypassing.

[14]Note that this way of looking at the data ignores any within-subjects effects.

Table 4: Multiple linear regression with order of presentation, Extreme Book FWMR scores and Fries BYPASS scores as predictor variables and Fries FWMR scores as the dependent variable

Model	B	SE	β	P
Constant	6.213	.526		<.0005
Order	-.128	.355	-.028	.719
Fries BYPASS	-.696	.095	-.593	<.0005
Extreme Book FWMR	.276	.088	.25	.002

Our results support this hypothesis: not only were FWMR scores in the deterministic case (Fries) highly negatively correlated with BYPASS scores, but they were better predicted by their BYPASS scores in the same deterministic case (Fries) than by their FWMR in a fatalistic case (Extreme Book). This suggests that participants are very sensitive to the sorts of conditions which compatibilists (and incompatibilists) think are relevant to whether free will and moral responsibility, and that participants are not merely automata attributing free will *no matter what*. This conclusion is strengthened by the fact that only 9% of participants both judged Extreme Book to involve bypassing (BYPASS > 4) *and* judged the protagonist to have free will (FWMR > 4). It is worth considering this figure in light of the fact that Feltz and Millan reported between 30 and 50% of participants having FWNMW intuitions across their various experiments.

5 Study 2: Free will no matter what?

If, as suggested by the results of Study 1, most participants have genuine compatibilist intuitions, rather than FWNMW intuitions, we should observe that most participants no longer attribute free will and moral responsibility to agents living in fatalistic universes once it is made clear that bypassing obtains in this universe. On the other hand, if participants really have FWNMW intuitions, then they should still attribute free will and moral responsibility, even if it is perfectly clear that agents' mental states are bypassed. In Study 2, we test these opposing predictions using a fatalist case in which clearly entails bypassing.[15] This procedure also allows us to reach a rough estimate of how many participants have FWNMW intuitions.

5.1 Participants

99 participants living in the United States were recruited using Amazon MTurk and paid $0.40 for their participation.[16] 6 were excluded for having incomplete answers.[17] 7 further were excluded for failing the comprehension question. Among the remaining 86 participants (51 male, 34 female, 1 no response), the mean age was 31.07 (SD=8.26).

5.2 Materials

All participants received the following case:

> More Extreme Book
> Imagine that there is a magic book that has all of our decisions and actions truly written in its content. For instance, whenever we are trying to decide what to do, the decision we end up making is completely and truly written in this book and the decision will happen regardless of our thoughts, beliefs, desires, and plans. The magic book has these events truly written in it lifetimes before the

[15] An anonymous referee notes that the wording of the case used, More Extreme Book, might also tap so-called 'deep-self' intuitions in its attempt to articulate strongly to participants the idea that the agent's mental states are bypassed as it includes the expression "deep down, John did not want to kill his wife". This might indeed be relevant and worth investigating further.

[16] One further participant entered the survey but didn't consent and progressed no further.

[17] We excluded those who didn't respond to each of the FWMR and BYPASS items.

> events took place. So, if the book has an event written in it, the event will definitely occur regardless of the past events and the laws of nature. Even if a person does not want to act this way, then she will be forced to act against her will by the book's magical powers. For example, one day a person named John decides to kill his wife so that he can marry his neighbor and he does it. Deep down, John did not want to kill his wife, and did not want to marry his neighbor. However, since these specific events were already and truly written in the book, it was impossible for John not to kill his wife regardless of his thoughts, beliefs, desires, and plans. Assume the book's contents made it impossible for John not to kill his wife. Please rate to what degree you agree with the following statements:

Participants then rated their agreement with the same 7 statements as in Study 1. Then, participants were asked a comprehension question and completed the Ten Item Personality Inventory. Finally, basic demographic information was gathered.

5.3 Results and Discussion

Mean answers and standard deviations for the 7 statements are reported in Table 5. The internal consistency among the FWMR items (Cronbach's α=0.8) and the BYPASS items (α=0.71) were a little lower than those observed in Study 1, but high enough to continue to calculate composite scores. FWMR scores for More Extreme Book (M=2.35, SD=1.6) were lower than for either Fries (M=4.62) or Extreme Book (M=3.27) in Study 1. Following the expected pattern, BYPASS scores (M=5.29, SD=1.48) were higher for More Extreme Book than either Fries (M=3.49) or Extreme Book (M=4.62) in Study 1.[18]

Table 5: Mean answers (and standard deviation) for each statement in Study 2.

	Statement	M	SD
1	John's/Bill's killing of his wife was up to him	1.9	1.58
2	John/Bill killed his wife of his own free will	1.93	1.66
3	John/Bill is morally responsible for killing his wife	3.22	2.35
4	John's/Bill's decision to kill his wife and children had no effect on what he ended up being caused to do	4.08	2.22
5	What John/Bill wanted had no effect on what he ended up being caused to do	5.59	2.09
6	What John/Bill believed had no effect on what he ended up being caused to do	5.77	1.85
7	John/Bill had no control over what he did.	5.71	1.91

How many participants continued to attribute free will and moral responsibility to the agent while interpreting his mental states as bypassed? To put it another way: how many did follow the "free-will-no-matter-what policy"? A nice way to answer this question is once again to conduct a midpoint split on FWMR and BYPASS scores (see Table 6).

Table 6: Midpoint split on FWMR and BYPASS scores in Study 2

			BYPASS	
			Disagree	Agree
FWMR	Extreme Book	Disagree	4.7%	73.3%
		Agree	7%	5.8%

As before, we treat values above/below the midpoint (4) as agree/disagree with the relevant idea (participants with scores of 4 are not included; percentages are of the total number of participants). Only 5 participants out of 86 (5.8%) agreed both that More Extreme Book involved bypassing and was compatible with free will and moral responsibility. Thus, it does not seem that so many participants have FWNMW intuitions. Even if one does not

[18] Neither FWMR nor BYPASS showed gender effects nor significant correlations with age or any of the five personality traits measured.

trust our way of measuring bypassing, for example because answers to BYPASS items might be influenced by answers to FWMR items (Rose and Nichols, 2013), and decides to take into account all participants who attributed free will and moral responsibility, the percentage of participants with FWNMW intuitions is at most 14%.

6 Conclusion

The results of Studies 1 and 2 strongly suggest the following: while some people do have FWNMW intuitions, it is a very small minority, and certainly not enough to support Feltz and Millan's error theory.

Feltz and Millan claim that many participants who apparently have compatibilist intuitions—who judge that people in a deterministic universe can act freely and be morally responsible for their actions—are, nonetheless, not *genuine* compatibilists. Indeed, according to Feltz and Millan, these apparent compatibilists would attribute free will and moral responsibility *no matter what*. Feltz and Millan support this theory by showing that many participants displaying apparent compatibilist intuitions also tend to say that people in a *fatalistic* universe can act freely and be morally—something a genuine compatibilist would supposedly not say.

We have shown that this conclusion is not warranted. Feltz and Millan's fatalistic cases are easily read as allowing agents to act on the basis of reasons and be the source of their actions—the cases do not necessarily imply *bypassing*. Under this reading of the cases, many compatibilists, e.g., source compatibilists, would be quite willing to say that one can act freely and be morally responsible for one's actions in such worlds, even if fatalism obtains. So, we hypothesized that Feltz and Millan's results were better explained by supposing not that many of their participants had FWNMW intuitions, but rather that many had broadly compatibilist intuitions – source compatibilist intuitions – and had read their fatalistic cases as allowing agents' mental states to be the source of their actions. We tested our hypothesis through two studies, the results of which supported our interpretation of Feltz and Millan's results. Study 1 showed that most participants who attribute free will and moral responsibility to agents in fatalistic universes do so mostly because they perceive these cases as implying bypassing. Study 2 showed that, once fatalistic scenarios are modified to clearly imply bypassing, very few participants (14%) maintained their judgment that an agent living in such a world can have free will and be morally responsible; moreover, of those, most interpreted the scenario as not involving bypassing. Thus, it seems that most participants displaying apparent compatibilist intuitions are indeed genuine compatibilists and that only a handful of them (5.8%) are ready to attribute free will no matter what.

Although our main concern was Feltz and Millan's claim that many apparent compatibilists are not genuine compatibilists, the implications of our findings are not limited to challenging Feltz and Millan's error-theory. First, our results bolster the case for the idea that most people are natural compatibilists. Attributions of free will and moral responsibility were above the midpoint for the Fries case, which clearly describes a deterministic universe. And second, and most importantly, our results provide insight into the *kind* of natural compatibilists they are. Our results suggest people are not leeway compatibilists so much as source compatibilists, since many participants are ready to attribute free will and moral responsibility to agents living in fatalistic universes, as long as these agents act on their own. In other words, what people seem sensitive to is an agent's ability to be the source of her actions and to act based on reasons, rather than her ability or inability to do otherwise.[19]

Acknowledgements

This research was supported by the National Center of Competence in Research (NCCR) Affective sciences financed by the Swiss National Science Foundation (n° 51NF40-104897) and hosted by the University of Geneva.

[19]There are reasons to be cautious about generalizing from participants recruited through Mturk. However, it has been observed that MTurk participants are slightly more demographically diverse than standard Internet samples, are significantly more diverse than typical American college samples, and that data obtained through Mechanical Turk are at least as reliable as those obtained via traditional methods (Buhrmester et al., 2011). Moreover, Study 1 replicated the results of Feltz and Millan's studies, showing that the results obtained through this kind of study can be reliably extended to other populations. Finally, it should be noted that intuitions about free will and moral responsibility seem to be robust and to display very little variation across cultures (Sarkissian et al., 2010).

References

Björnsson, G. and Pereboom, D. (2014). Free will skepticism and bypassing, in W. Sinnott-Armstrong (ed.), *Moral Psychology*, Vol. 4, MIT Press, pp. 27–35.

Buhrmester, M., Kwang, T. and Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?, *Perspectives on Psychological Science* 6: 3–5.

Cova, F. (2014). Frankfurt-style cases user manual: Why frankfurt-style enabling cases do not necessitate tech support., *Ethical Theory and Moral Practice* .

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

Cova, F. and Kitano, Y. (forthcoming). Experimental philosophy and the compatibility of free will and determinism: a survey, *Annals of the Japan Association for Philosophy of Science* .

Diderot, D. (1986). *Jacques the fatalist*, Penguin.

Feltz, A. (2009). Experimental philosophy, *Analyze and Kritik* 31: 201–219.

Feltz, A. and Cova, F. (2014). Moral responsibility and free will: A meta-analysis, *Consciousness & Cognition* 30.

Feltz, A. and Millan, M. (forthcoming). An error theory for compatibilist intuitions, *Philosophical Psychology* .

Fischer, J. M. (1986). Responsibility and failure, *Proceedings of the Aristotelian Society* 86: 251–270.

Fischer, J. M. and Ravizza, M. (1991). Responsibility and inevitability, *Ethics* 101: 258–278.

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

Gosling, S. D., Rentfrow, P. J. and Swann, W. B. (2003). A very brief measure of the big-five personality domains, *Journal of Research in Personality* 37: 504–528.

Kane, R. (2005). *A Contemporary Introduction to Free Will*, OUP.

Miller, J. S. and Feltz, A. (2011). Frankfurt and the folk: An experimental investigation of frankfurt-style cases, *Consciousness & Cognition* .

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

Nahmias, E., Morris, S. G., Nadelhoffer, T. and Turner, J. (2006). Is incompatibilism intuitive?, *Philosophy and Phenomenological Research* .

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

Pereboom, D. (1995). Determinism al dente, *Noûs* .

Ravizza, M. (1994). Semi-compatibilism and the transfer of non-responsibility, *Philosophical Studies* .

Rose, D. and Nichols, S. (2013). The lesson of bypassing, *Review of Philosophy and Psychology* .

Roskies, A. and Nichols, S. (2008). Bringing moral responsiblity down to earth, *Journal of Philosophy* .

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S. and Sirker, S. (2010). Is belief in free will a cultural universal?, *Mind & Language* 25(3): 346–358.

Sommers, T. (2010). Experimental philosophy and free will, *Philosophy Compass* .

Strawson, P. F. (1962). Freedom and resentment., *Proceedings of the British Academy* .

Van Inwagen, P. (1983). *An Essay on Free Will*, OUP.

Weigel, C. (2013). Experimental evidence for free will revisionism, *Philosophical Explorations* .