A Defense of Natural Compatibilism

Florian Cova

Philosophy Department, University of Geneva

Penultimate draft

Final version to appear in J. Campbell, K. M. Mickelson and V. A. White (Eds.), *Blackwell Companion to Free Will*. Blackwell.

1. Investigating folk conceptions of free will and moral responsibility

If we define FREE WILL as the type of control one has to exert upon one's action to be morally responsible for them (see for example Mele, 2008)[1], then FREE WILL is not a technical concept coined by philosophers, such as ALIEFS, EMERGENCE, or TROPES. Rather, such a concept seems to be already present in our everyday practices – at least at an implicit level, where it guides our assessment of others' moral responsibility, the reactive attitudes we adopt towards them and certain behaviors such as rewards and punishment. Long before Aquinas introduced the expression *liberum arbitrium*, people were distinguishing between people who acted willfully and those who acted under duress, and responded to their actions accordingly.

Thus, it is not a stretch to say that philosophical reflection on free will takes its roots in an intuitive notion that pervades our interactions with others. Recently, philosophers and psychologists have tried to get a better understanding of this everyday understanding of free will. Their research has focused on two different questions. The first one is *what kind of control do people think we need to have over our actions to be morally responsible for them*?

[1] Of course, not every philosopher accepts this definition of free will (see for example van Inwagen, 2008). However, for reasons that will get clearer as we progress, this is the one I will be using here. One objection to this kind of definition is that "moral responsibility" comes in different varieties (Rossi & Warfield, 2017). In the context of this paper, I will use "moral responsibility" in the sense of *accountability* (Watson, 1996).

The second is *what kind of control do people think we actually have on our actions*? In this chapter, I will focus on the first question, as it is the one that has generated the most research and debate in the past years (though I will briefly touch on the second towards the end of the chapter).

1.1. The relevance of folk intuitions to the philosophical debate on free will

Besides the obvious psychological importance of these questions for our understanding of the human mind and social cognition, one might wonder what is the relevance of such inquiries for philosophical theorizing about free will. An answer is that intuitions about the conditions required for moral responsibility play a major role in philosophical debates about the *nature* of free will: most arguments take as their premise either principles that are supposed to be self-evident (e.g. the Principle of Alternate Possibilities, van Inwagen's Rule Beta) or intuitions about individual thought-experiments (e.g. Frankfurt cases, Manipulation arguments).

However, despite some attempts (e.g. Double, 1996), the free will debate is still in need of an explicitly articulated meta-philosophy, and it is not clear what the status and role of such intuitions are. There are several ways to understand the role intuitions play in the philosophical debate about the nature of free will. According to what I call the *weak view*, intuitions only set the default point and determine on who lies the burden of proof (e.g. Nahmias et al., 2006). According to the *moderate view*, intuitions give us access to certain modal truths about free will and moral responsibility (e.g. that it is *possible* for someone to be morally responsible even if one could not have done otherwise) and constitute *prima facie* defeasible evidence in their favor. Finally, according to what I call the *strong view*, intuitions have the power to determine the subject matter of the free will debate, either because widespread intuitions and truisms about free will and moral responsibility actually determine the reference of the expression ‘free will’, or because these intuitions provide the main basis for a conceptual analysis of our shared concept of FREE WILL, in a way that makes it impossible for an adequate theory of free will to ignore our intuitions about it (e.g. Jackson, 1998). Whatever your position on these matters, it seems clear that getting a better grasp of the intuitive ground from which the philosophical problem of free will emerged might prove useful in evaluating certain philosophical arguments.

1.2. Two main positions: natural compatibilism vs. natural incompatibilism

In their study of folk intuitions about the *nature* of free will, experimental philosophers and psychologists have mainly focused on the *compatibility question*: do people consider determinism to prevent the possibility of free will? *Natural Incompatibilism* is the claim that laypeople conceive free will in such a way that it is incompatible with determinism, because they take determinism to be an obstacle to free will. *Natural Compatibilism* is the claim that free will as we intuitively think of it is perfectly compatible with determinism. Though some have argued that the very same people can sometimes have compatibilist and incompatibilist intuitions depending on the context (Cova, 2011; Doris, Knobe & Woolfolk, 2007; Knobe & Nichols, 2010), most of the debate in the literature has been framed as an opposition between *Natural Compatibilism* and *Natural Incompatibilism*. In this chapter, I will review this experimental literature and argue that, if we were to choose between the two, *Natural Compatibilism* would be a better fit of the available data.

But before we start, I need make two additional remarks. The first one is about the material I will cover in this chapter. As pointed out by Cova & Kitano (2014) and Feltz (2017), these studies come in two kinds. A first group set of studies tries to *directly* address the compatibility question, by investigating whether people think that an agent in a deterministic universe can be free and morally responsible. A second group *indirectly* addresses the question by studying whether people share the intuitions that serve as premise to philosophical arguments for and against the compatibility of free will with determinism, such as Frankfurt-style cases (Cova, 2014, 2017; Miller & Feltz, 2011) or Manipulation arguments (Björnsson, 2016; Cova, forthcoming; Feltz, 2013; Sripada, 2012). In this chapter, I focus on studies that try to *directly* address the compatibility question.

The second remark is about the way these studies have operationalized and measured free will. Some studies directly ask participants about free will (e.g. by asking them whether the agent *freely* did a certain thing, or did a certain thing *of his own free will*), other ask participants about moral responsibility (e.g. by asking whether the agent is *morally responsible* for doing something, or deserves *blame/praise* or a *punishment/reward* for what they did), and others ask both types of questions. As mentioned at the start of this chapter, I will focus on free will defined as the type of control required by moral responsibility – meaning that, when both kinds of measure will yield different conclusions, I will take participants' answers about moral responsibility as the most reliable indicator. Indeed, though investigating how participants use terms such as 'free will' or 'freely' might be interesting in its own right, it is not clear how these are relevant to the philosophical debate, as very few philosophical arguments seem to depend on linguistic intuitions about the way such words are used in everyday language. Moreover, while practices involving attributions of moral responsibility (such as reactive attitudes or punishment) seem pervasive and human universals, some cultures seem to lack words or expressions that would be equivalent to 'free will' (Berniūnas et al., 2021). Finally, most people take philosophical debates about free will to be important precisely because of their practical relevance and the implications they have for our image of ourselves as morally responsible agents. For all these reasons, I will thus focus on participants' intuitions about moral responsibility when possible.

2. A short history of the Natural Compatibilism vs. Natural Incompatibilism debate

2.1. Folks as natural compatibilists

The first studies on the compatibility question were published in 2005 and 2006 by Nahmias and his colleagues (Nahmias et al., 2005, 2006). The principle of these studies was quite simple: participants were presented with a description of a deterministic universe, then were told about an agent living in this universe performing a particular action. Participants were then asked whether the agent was morally responsible for this action, and acted of his own free will.

Here is an example:

> SUPERCOMPUTER – 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 participants answered that Jeremy acted of his own free will. To the second, 83% of participants answered that Jeremy was morally blameworthy. In another version, Jeremy did not rob a bank but saved a child from a burning building. To this scenario, 68% of participants answered that Jeremy saved the child of his own free will and 88% 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 participants answered that Jeremy went jogging of his own free will.

Nahmias and his colleagues obtained similar results for scenarios describing a universe in which one's actions and decisions are fully determined by the combination of one's genes and upbringing, and for scenarios describing a universe submitted to *Eternal Recurrence*, where the same things are doomed to happen again and again. Together, these results suggest that *most* (though not all) of their participants have compatibilist intuitions: they judge free will and moral responsibility to be compatible with determinism (see Figure 1).

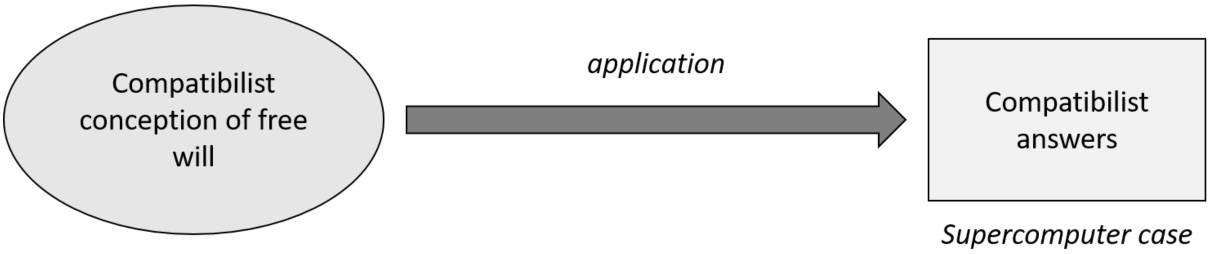

Figure 1. Folks as natural compatibilists.

2.2. Folks as natural incompatibilists: the Performance Error Model

However, things might not be so straightforward. Noticing that the scenarios used by Nahmias and his colleagues all make use of ‘concrete’ cases, in which participants are asked about a single action performed by a given, identifiable individual, Nichols and Knobe (2007) investigated how participants would respond to more ‘abstract’ questions. They first presented participants with the following description of two 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.

Obviously, Universe A is intended to be a deterministic universe, while Universe B, by contract, is intended to be an indeterministic universe. After reading this description, participants were randomly assigned to the *concrete* or *abstract* condition. Participants in the *concrete* condition received the following vignette:

> 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 participants (72%) gave the compatibilist answer according to which the agent was fully morally responsible, even if he lived in deterministic universe A. These results are consistent with those obtained by Nahmias and his colleagues. But, let’s now consider 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 participants (86%) gave the *incompatibilist* answer. To put it otherwise: participants tended to deny that agents could be morally responsible for their action in the deterministic universe A. Thus, participants' answers to experimental philosophers' probes seem incoherent. How are we to explain these results?

Nichols and Knobe have their own explanation; the Performance Error Model. According to this model, the folk conception of free will is, at its core, perfectly incompatibilist: free will, as we naturally conceive it, is incompatible with determinism. This commitment is what is reflected in the abstract condition, when participants are not asked to judge and condemn a particular individual: they give mostly *incompatibilist* answers. However, when faced with a concrete case (that is: a case featuring a given agent performing a precise action), participants' intuitions and judgments can be swayed by their emotions. Indeed, particular actions can be revolting and outrageous (as Billy's murder of his wife and three children), thus eliciting the desire to punish the agent. This desire, in turn, leads people to justify their desire to punish the agent by attributing him a certain amount of moral responsibility (Figure 2).

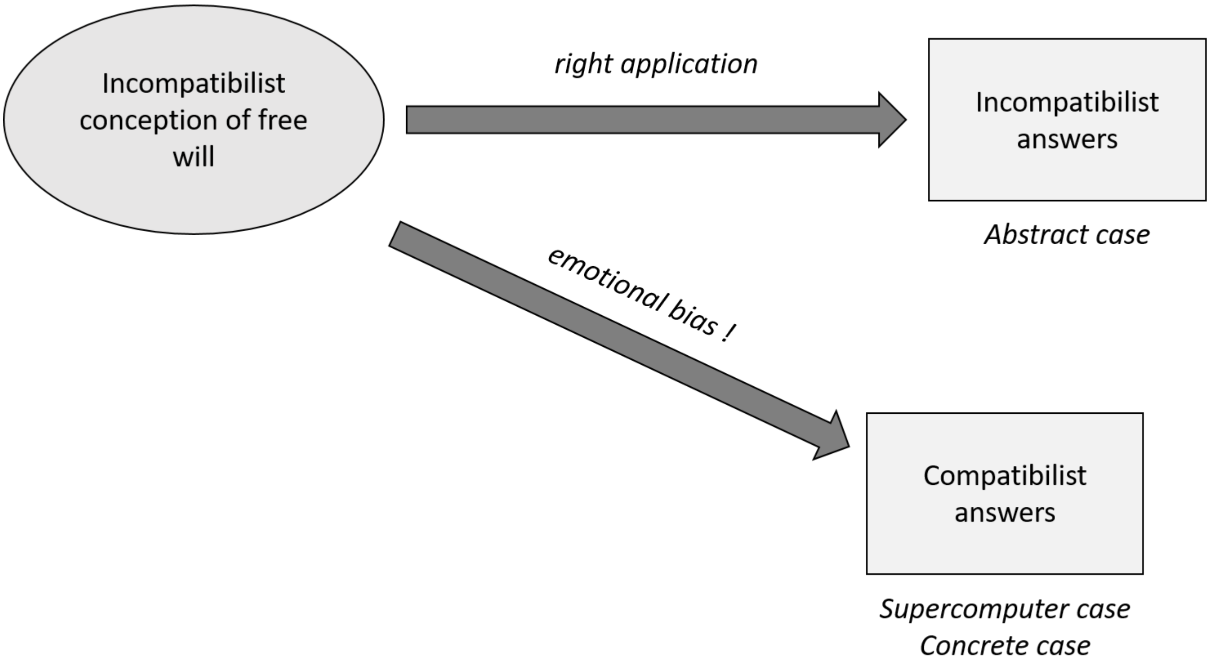

Figure 2. Nichols and Knobe's Performance Error Model.

Nichols and Knobe's Performance Error Model is in line with other findings in social psychology, namely that people are more likely to believe in free will when they have just been presented with descriptions of moral violations (Clark et al., 2014). It is also supported by additional evidence presented by Nichols and Knobe themselves. Indeed, one prediction of the Performance Error Model is that participants should give mostly *incompatibilist* answer when presented with a concrete case that is not emotionally salient. To test this prediction, 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?

Participants were assigned either to the *low affect* or *high affect* condition, and told either that the agent lived in the (deterministic) Universe A or that he lived in the (indeterministic) Universe B. Table 1 describes, for each case, the proportion of participants who judged that the agent could be fully morally responsible for his action.

	Agent in deterministic universe A	Agent in indeterministic universe B

High Affect	64%	95%
Low Affect	23%	89%

Table 1. Results from Nichols & Knobe (2007)

As predicted by Nichols and Knobe’s Performance Error Model, participants in the *low affect* condition tended to judge that the agent could not be responsible for cheating on his taxes when he lived in a deterministic universe. On the contrary, participants in the *high affect* condition judged that the agent was morally responsible for raping his victim, even when he lived in a deterministic universe. Thus, it seems that participants’ intuitions perfectly fit the Performance Error Model.

However, there is a problem with Nichols and Knobe’s results: other experimental philosophers have had a lot of trouble replicating them. Indeed, though the difference between the *abstract* and *concrete* condition seems robust (and has been replicated in several countries, see Sarkissian et al., 2010), other experimental philosophers had trouble replicating the difference between the *low affect* and *high affect* conditions. A meta-analysis by Feltz and Cova (2014), which aggregated all known attempts at replicating this effect, found that the real difference between the *low* and *high* affect conditions was actually very small – and thus that the effect of emotional reactions on judgments about moral responsibility were not strong enough to actually explain the huge difference between the *abstract* and *concrete* condition.

Another problem for the Performance Error Model comes from a study conducted by Cova, Bertoux and their colleagues (2012) on patients suffering from a behavioral variant of frontotemporal dementia, a neurodegenerative disease accompanied by a deficit in emotional responses. Given bvFTD patients’ impaired emotional reactions to descriptions of gruesome and violent acts, the Performance Error Model should predict that bvFTD patients, being free of emotional biases, should give more incompatibilist answers than control participants to concrete cases. Thus, Cova and his colleagues presented participants with the *supercomputer* and *concrete* cases. However, bvFTD patients gave the same answers as control participants – that is: mostly compatibilist answers.

Thus, it seems that participants' compatibilist answers cannot be explained away as the mere effect of emotional biases. We thus need another explanation for the apparent incoherence in participants' judgments about moral responsibility.

2.3. Folks as natural compatibilists: the Bypassing Hypothesis

This other explanation is Murray and Nahmias' Bypassing Hypothesis (Murray & Nahmias, 2014). As the Performance Error Model, the Bypassing Hypothesis was originally designed to account for an apparent incoherence in participants' judgments. Indeed, as a follow-up on his prior investigations, Nahmias (2006) decided to investigate whether the *kind* of determinism participants were presented with would make a difference (see also Nahmias, Coates & Kvaran, 2007). This is why he 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 participants answered positively to the first question and 77% answered positively to the second question, thus giving perfectly compatibilist answers. In the *neurological determinism* condition, only 18% of participants answered positively to the first question and 19% to the second question, that is: clearly incompatibilist answers. To put it otherwise: participants gave mostly compatibilist answers in the first case and mostly incompatibilist answers in the second case. Once again, there seems to be something incoherent in participants' judgments about free will and moral responsibility.

However, Nahmias offers a plausible explanation for this pattern of judgment. According to him, this asymmetry arises because people take neurological determinism, but not psychological determinism, to imply that people's mental states play no role in the generation of their decisions and actions – that people's mental state are, so to speak, *bypassed*. But, both compatibilists and incompatibilists alike would agree that we would not be free nor morally responsible for our actions if our desires, beliefs and values played no role in our decisions and actions. Thus, participants' seemingly 'incompatibilist' answers to *neurological determinism* case do not show that they have an incompatibilist understanding of free will – only that (i) they take neurological determinism to imply 'bypassing' and (ii) that they take 'bypassing' to exclude free will and moral responsibility. On the other hand, their compatibilist answers to the *psychological determinism* case reflect their true, compatibilist understanding of free will.

According to Murray and Nahmias (2014, see also Nahmias and Murray, 2011), the same kind of explanation can be applied for the difference between Nichols and Knobe's *abstract* and *concrete* conditions. For them, Nichols and Knobe's description of determinism is misleading and lead participants to assume not only that Universe A is deterministic (i.e.

that every human action in it is fully caused by prior events), but also that it involves ‘bypassing’ (i.e. that agents’ mental states do not play a causal role in the generation of their actions and decisions). This is why participants who are presented with this case and asked the *abstract* question give seemingly ‘incompatibilist’ answers: they are just expressing their belief that moral responsibility is incompatible with ‘bypassing’.

But why do participants tend to give compatibilist answer in the *concrete* case, then? Here, the idea is that, because the concrete case clearly states that the agent acts on the basis of his mental states difference (he kills his wife and children because he *wants* to be with his secretary), participants might be less likely to infer that determinism entails bypassing in these cases. Hence the mostly compatibilist answers (see Figure 3).

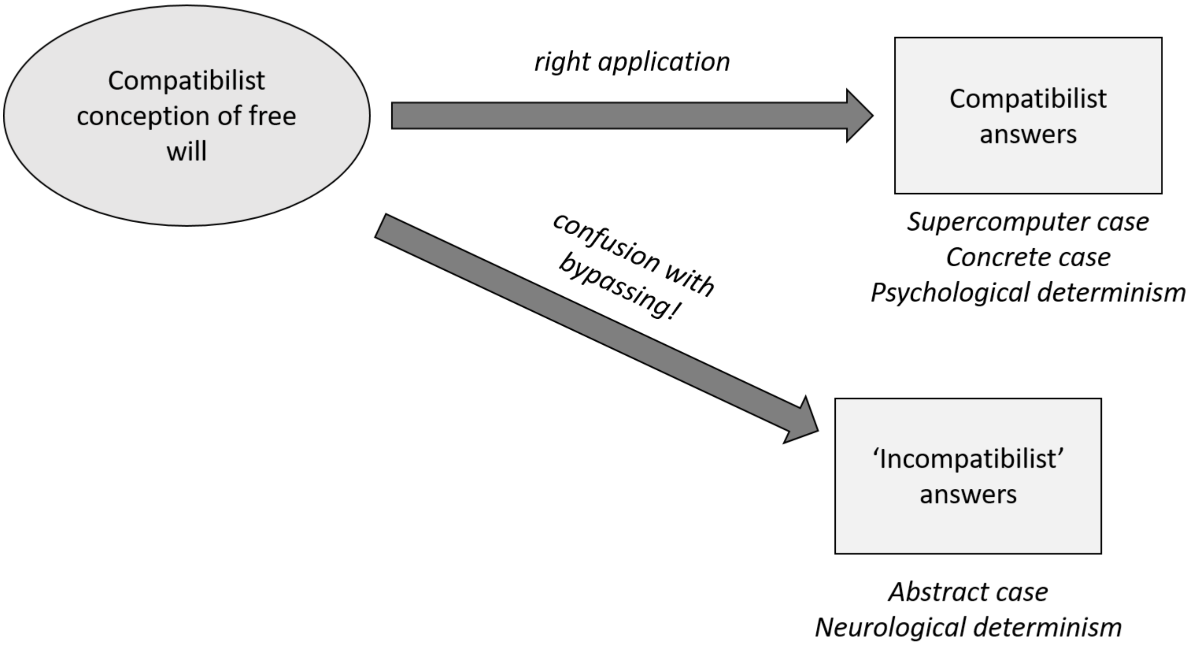
Compatibilist conception of free will
right application
Compatibilist answers
Supercomputer case
Concrete case
Psychological determinism
confusion with bypassing!
‘Incompatibilist’ answers
Abstract case
Neurological determinism

Figure 3. Nahmias and Murray’s Bypassing Hypothesis.

To test these predictions, Nahmias and Murray (2011) presented participants either with Nichols and Knobe’s *abstract* and *concrete* condition (in Universe A), or with an abstract and a concrete version of the eternal recurrence case used by Nahmias and his colleagues (2006). For each scenario, 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, and to which extent they were likely to confuse determinism with bypassing:

> **Decisions:** In Universe [A/C], a person's decisions have no effect on what they end up being caused to do.
>
> **Wants:** In Universe [A/C], what a person wants has no effect on what they end up being caused to do.
>
> **Believes:** In Universe [A/C], what a person believes has no effect on what they end up being caused to do.
>
> **No Control:** In Universe [A/C], a person has no control over what they do.

Participants' answers to these questions were aggregated to constitute a *bypassing* score. The higher the score, the more participants were likely to confuse determinism with bypassing. Then, Nichols and Murray conducted a *mediation analysis* to determine to which extent participants' tendency to confuse determinism with bypassing explained the impact of different scenarios on participants' judgments about free will and moral responsibility. The results of this analysis (presented in Figure 4) revealed that bypassing scores *fully mediated* the effect of scenarios on judgments of free will and moral responsibility, suggesting that participants' incompatibilist judgments are indeed explained by their taking determinism to imply bypassing. To put it otherwise: seemingly incompatibilist judgments seem to be the product of a confusion, and when people understand exactly what determinism is, they tend to give compatibilist answers.

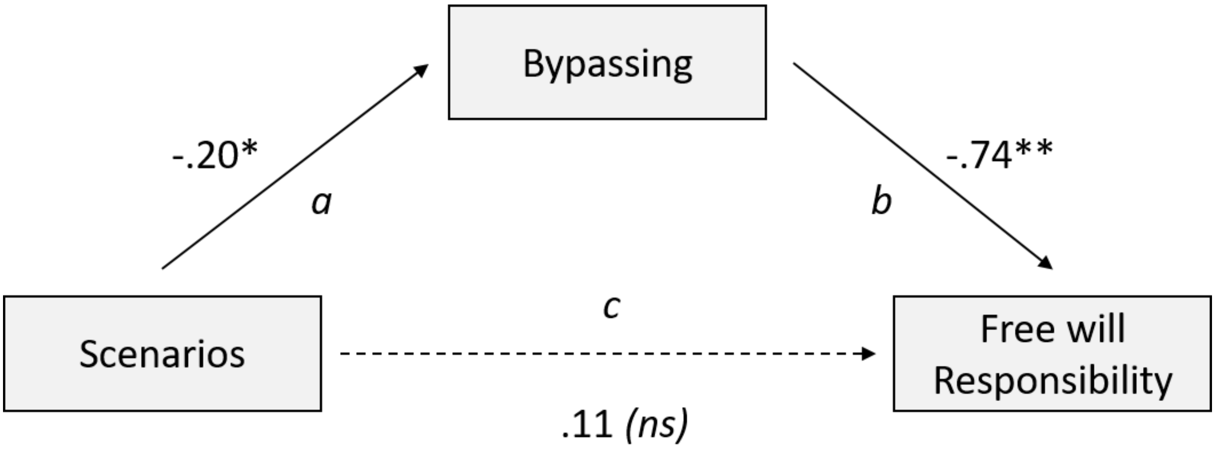

Figure 4. Mediation analysis by Nahmias and Murray (2011)

3. Answering methodological objections to the Bypassing Hypothesis

3.1. Two objections to the Bypassing Hypothesis

The Bypassing Hypothesis seem to support the idea that people are mostly compatibilists at heart (what we called *Natural compatibilism*). However, for all its merits, the Bypassing Hypothesis has drawn several criticisms.

A first set of criticisms comes from Rose and Nichols (2013). They argue that Nahmias and Murray's model is not the only (nor the best) explanation for their results. Indeed, it turns out that, re-analyzing Nahmias and Murray's results, one can find that free will and moral responsibility scores (henceforth: FW/MR scores) also mediate the effect of scenarios on bypassing scores. So, Nahmias and Murray's data are compatible with two different models:

- The Bypassing Hypothesis: Scenarios → Bypassing scores → FW/MR scores
- The Incompatibilist Model: Scenarios → FW/MR scores → Bypassing scores

To determine which of these two models is the best, Rose and Nichols recruited participants that were presented with Nichols and Knobe's *abstract* condition, either situated in Universe

A or Universe B. Participants were asked questions about free will, moral responsibility, and bypassing. Rose and Nichols then used structural equation modelling and the TETRAD IV program to compare the Bypassing Hypothesis and the Incompatibilist Model. They found that the Incompatibilist Model was a better fit of the data than the Bypassing Hypothesis (see Figure 5).

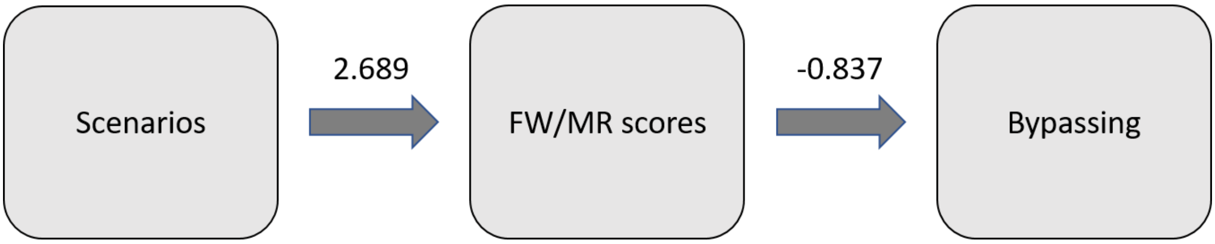

Figure 5. Rose and Nichols' Incompatibilist Model. Arrows represent a direct, causal connection. The numbers above each edge are linear coefficients

A second, related set of criticisms was put forward by Björnsson (2014). Björnsson ran a new study using questions about free will, moral responsibility and bypassing, but also introducing a new type of question, which he calls *throughpass questions*:

> **Abstract Throughpass:** In Universe A, when earlier events cause an agent's action, they typically do so by affecting what the agent believes and wants, which in turn causes the agent to act in a certain way.
>
> **Concrete Throughpass:** When earlier events caused Bill's action, they did so by affecting what he believed and wanted, which in turn caused him to act in a certain way.

Throughpass questions are supposed to be the polar opposite of bypassing questions: they present a universe in which agents' mental states play a crucial role in the generation of their action. Thus, if bypassing questions are interpreted by participants as intended by Nahmias and Murray, we should observe an inverse correlation between participants' agreement to the throughpass and bypassing statements.

Björnsson's results raise two problems for the Bypassing Hypothesis. First, he found that bypassing scores did not fully mediate the effect of scenarios on FW/MR scores, which suggests that participants' tendency to confuse determinism with bypassing does not completely explain their incompatibilist judgments. Second, he found a weak but *positive* correlation between bypassing and throughpass statements, which suggests that participants might not interpret bypassing questions as intended by Nahmias and Murray.

Taken together, these findings cast doubt on the validity of the Bypassing Hypothesis. Should we then discard the Bypassing Hypothesis? I don't think so. Why? Here is the short answer: because, even it faces some issues, it still constitutes a very good explanation for a range of phenomena, such as the difference between the *abstract* and *concrete* condition and the difference between psychological and neurological determinism. Rose and Nichols' Incompatibilist Model might well be a better fit of the data, it does not have this kind of explanatory power. In fact, it is not even clear whether there is a good psychological explanation that would correspond to the Incompatibilist Model. Why would judgments about free will and moral responsibility drive judgments about bypassing? Is it because participants first judge that free will and moral responsibility are impossible in a deterministic universe, then try to justify this judgment by appealing to bypassing-related considerations? But if participants were truly incompatibilists (and not compatibilists), they should not feel the need to point at such considerations to justify their judgment. Thus, it is not clear how we are supposed to make psychological sense of the Incompatibilist Model and, despite its problems, the Bypass Hypothesis is still the best available explanation for participants' judgments in a wider range of cases.

3.2. Answering the Throughpass objection

Now, for those who would not be satisfied with this short answer, there is a longer, much more technical one. Let’s first begin with Björnsson’s most puzzling finding: that there is a *positive* correlation between bypassing and throughpass statements, that are supposed to measure opposed constructs. How are we to account for these results?

To explain them, I must introduce a distinction between the *deep* self (and deep motivational states) and the *superficial* self (and superficial motivational states). In his 1889 book on *Time and Free Will,* French philosopher Henri Bergson developed a conception of the human mind as composed of several layers (Bergson, 1889/2002). The mental states that belong to the innermost layer constitute the ‘deep self’: these are the values and commitments that define who we *really* are. The outermost, shallow layers are constituted by mental states that are not truly ‘ours’: habits, automatism, social conditioning and pressure, etc. According to Bergson, to be free is to break the ‘crust’ of shallow mental states to act on the basis of our ‘deep self’.

Similar ideas can be found in contemporary analytical philosophy (Watson, 1975; Wolf, 1993; Sripada, 2016). More importantly, recent empirical studies have confirmed that most people naturally think this way, and distinguish between one’s deep self and one’s superficial self (Cova, 2011; Newman, Bloom & Knobe, 2014, Sripada, 2012). Thus, people naturally make a distinction between two kinds of mental states: those who belong to the deep self, and reflect what we really care about, and those who belong to the superficial self, and can go against what we *really* believe and want.

With this distinction at hand, we have an easy explanation for Björnsson’s results. Take the following example: John really loves Sally. But Black, a master hypnotist hypnotizes John to induce in him the irrepressible desire to be mean and cruel to Sally. Now, John has two kinds of mental states: a desire to help and protect Sally (that reflect what he really cares about, and is thus part of his deep self) and a desire to harm her (that has been induced by Black, and is thus part of his superficial self). Now, imagine that we present participants with this case. We then ask them how much they agree with the corresponding throughpass statement:

> When earlier events caused John’s action, they did so by affecting what he believed and wanted, which in turn caused him to act in a certain way.

Here, we can expect most participants to agree with this statement: after all, when Black causes John to be mean to Sally, he does so by affecting John’s mental states (i.e. desires). But, let’s now imagine that participants are presented with the following bypassing statement:

> What John wants has no effect on what he ends up being caused to do.

Would participants agree with this statement? In one sense, they should disagree: if John is now mean to Sally, it is precisely because he wants to. But, at the same time, there is a sense in which this statement is true: what John *really* wants (i.e. being kind to Sally) has actually no effect of what he ends up being caused to do. The statement is thus ambiguous: should “what John wants” be interpreted in a broad way, as encompassing his superficial desires and wants, or in a more restricted sense, as “what he *really* wants”?

Let’s now imagine that the majority of participants actually adopt the narrow interpretation, and think they are asked about what John *really* wants. In this case, we should expect agreement with the bypassing statement to be high. And, thus, we obtain a positive correlation between throughpass and bypassing statements. But this correlation does not constitute a contradiction: one can think that John’s actions are caused by his mental states, without accepting that they are caused by his *deep* mental states. The same is true for

Björnsson's results: if participants take determinism to imply that one is caused to act only by their superficial mental states, and thus that deep mental states play no role in the generation of one's actions, we should also observe the positive correlation Björnsson observed.

Is this explanation of Björnsson's results the right one? To find out, I conducted an online study in which participants were presented with one vignette introducing two characters: John, a gifted and ambitious neuroscientist who needs money, and Bill, the neighbor of John's aunt. At the end of the vignette, John always made it so that Bill killed the aunt and that John inherited her money. The vignette existed in 6 different version, varying across two dimensions. The first dimension was *Manipulation*: how John made it so that Bill ended up killing the aunt. In one case (*Body control*), John directly took control of Bill's body, not intervening on Bill's mental states. In another case (*Mind control*), John induced in Bill a strong and irresistible urge to kill John's aunt. Finally, in the third option (*Money*), John, knowing that Bill was in desperate need of money, simply offered Bill money in exchange of killing his aunt. The second dimension was *Reasons*: whether Bill had personal reasons *for* killing John's aunt (he hated her) or had personal reasons *against* killing her (they were friends). After reading the vignette, participants were asked to indicate their agreement (on a scale from -3 to 3) with a series of statements, including (i) statements about Bill's moral responsibility, statements about Bill's free will, (iii) bypassing statements, and (iv) a throughpass statement. The results of this study are presented in Table 2.

	Body control		*Mind control*		*Money*	
Reasons	For	Against	For	Against	For	Against
Resp.	-2.15 (1.26)	-2.38 (1.28)	-1.08 (1.81)	-1.94 (1.32)	2.12 (1.09)	2.21 (0.99)
Free Will	-2.70 (0.94)	-2.77 (1.05)	-2.16 (1.52)	-2.74 (0.70)	0.58 (2.10)	0.55 (2.06)
Bypass	1.85 (1.38)	2.39 (1.07)	1.13 (1.46)	1.89 (1.29)	-1.44 (1.05)	-1.27 (1.15)

Throughpass	0.11 (2.13)	0.59 (2.39)	1.08 (1.65)	1.43 (1.88)	0.59 (1.59)	0.78 (1.53)
N	87	73	75	90	83	91

Table 2. Mean and SDs for participants’ answers to the manipulation study in function of type of *Manipulation* and *Reasons* for responsibility, bypassing and throughpass scores.

As expected, bypassing scores were very high for the *Body control* condition (in which the agents’ mental states played no role) and low in the *Money* condition (in which the agent acts on his own). Bypassing scores in the *Mind control* condition (in which the agent acts on the basis of a mental state, but one that is externally induced) were significantly lower than in the *Body control* condition, but still higher than midpoint, and way higher than in the *Money* condition. This shows that most (but not all) participants interpret the bypassing statements as not asking about any kind of mental state, but about *deep* and/or *authentic* mental states.

Participants’ ratings for the throughpass statements were puzzling. Even in the *Body control* condition, participants tended to agree with the throughpass statement. Moreover, compared to the *Body control* condition, throughpass ratings were not significantly higher in the *Money* condition. Even worse, their agreement with the throughpass statement was significantly higher in the *Mind control* condition than in the *Money* condition. As predicted, this suggests that the throughpass statement absolutely fails to measure what it was designed to measure (whether an agent’s mental states play a crucial role in the formation of their action) and rather seems to measure to which extent the agent is acted by external forces. Thus, defenders of the Bypassing Hypothesis do not need to worry about the fact that throughpass and bypassing statements are not negatively correlated.

Before we move on, one interesting result should be noted: though above the midpoint, attributions of free will were quite low in the *Money* condition, compared to moral responsibility attributions. This is probably due to the fact that external circumstances (dire lack of money) put pressure on Bill to accept John's offer. Thus, some participants seem to use 'free will' in a very demanding sense, in which one has to make decision independently from external pressures to count as 'free'. However, this sense is probably not the one relevant to theoretical theorizing, as most philosophical accounts of free will would certainly conclude that Bill killed John's aunt of his own free will. Thus, this is one more argument in favor of the conclusion that moral responsibility ratings might do a better job at tracking the philosophically relevant concept of free will.

3.3. Answering the 'best-fit-of-the-data' objection

Now, what about the fact that bypassing statements do not fully mediate the effect of scenarios on FW/MR judgments? My reply here is going to be technical but can be summarized as follows: this objection presupposes that, if participants' perceptions of bypassing explain the effect of determinism on their free will and moral responsibility judgments, then we should expect participants' agreement with bypassing statements to *fully* mediate the effect of determinism on their free will and moral responsibility judgments. However, this assumption is simply wrong.

More precisely, this assumption would be right if we were able to measure with complete accuracy participants' perceptions of bypassing, free will and moral responsibility. But this is an unreasonable assumption, as items measures are susceptible to be interpreted in different ways by different participants. And, as soon as some measurement error is introduced, expectations of full mediations might no longer be reasonable.

To make this answer more salient, I decided to simulate some data. I distributed 200 virtual participants across two conditions: *Determinism* and *Indeterminism* (100 in each condition). Participants in the *Determinism* condition were randomly assigned a Bypassing score between 1 and 4, and those in the *Indeterminism* condition were randomly assigned a score between 4 and 7. Free Will scores were randomly computed from Bypassing scores by adding -1, 0 or 1 to Bypassing scores. Thus, we simulate a causal chain such as: condition → bypassing → free will.

When we analyze these simulated data, we find as expected that the effect of condition on Free Will is fully mediated by Bypassing, but that the effect of condition on Bypassing is *not* fully mediated by Free Will. So, the results of mediation analysis accurately describe the underlying causal structure.

However, in this case, our measures *perfectly* represent the underlying phenomena. What happens when a little noise is added? To find out, I randomly added some noise to the data by randomly adding -1, -0.5, 0, 0.5, 1 to the Bypassing and Free Will scores. The new 'inaccurate' scores were still strongly correlated with the original scores (r = .92 and .93 respectively), meaning that only a little measurement error was introduced. But this was already enough to make it so that Bypassing no longer *fully* (but only *partially*) mediated the effect of condition on Free Will (see Figure 6).

So much for the full mediation criterion, but what about choosing the model that best fits the data? To find out, I added more noise in Bypassing scores to simulate a situation in which the measurement error for Bypassing scores is greater than for Free Will scores (so that the correlation between original and actual Bypassing scores was 0.62). I then entered these data in the TETRAD program and asked the Greedy Search Algorithm to search for the model that best fitted the data. As can be seen in Figure 6, the model selected by the algorithm did not match the original causal structure. I also asked TETRAD to compare two causal models of the data: (i) condition → bypassing → free will, and (ii) (i) condition → free will → bypassing. It concluded that the second model was a way better fit than the first one. Thus, measurement error also compromises the use of such methods to determine which causal structure is the right one, particularly when the measurement error is higher for the mediator than for the variable we seek to predict.

Are we in such a situation? I think we can reasonably say ‘yes’. Participants’ judgments about free will and moral responsibility are precisely the outcome we seek to measure so, barring lies and inattention, their answer should accurately reflect the phenomenon we are interested in. But we saw that bypassing statements were subject to ambiguity: some participants took them as speaking of agents’ mental states in general (including externally induced mental states) while other interpreted them as asking about the agent’s *deep* and *authentic* mental states. This ambiguity naturally translates in measurement error. Thus, we are clearly in a situation where giving too much weight to the fact that there is not a *full* mediation or that a model best fit the data than another might lead us to make wrong theoretical choices.

Not that I am rejecting the use of mediation analysis as a *test* for a certain hypothesis: clearly, if a theoretical account predicts that a phenomenon M explains the connection between two other phenomena A and B, we should expect measures of M to mediate the link between measures of A and B. However, present measurement error, it might not be reasonable to demand that this mediation be a *full* mediation or this model to be the best fit of the data, as measurement error tends to underestimate the indirect effect and overestimate the direct effect (VanderWeele, Valeri & Ogburn, 2012). Such statistical considerations should not trump other arguments for the theoretical model, such as its ability to explain a whole range of phenomena.

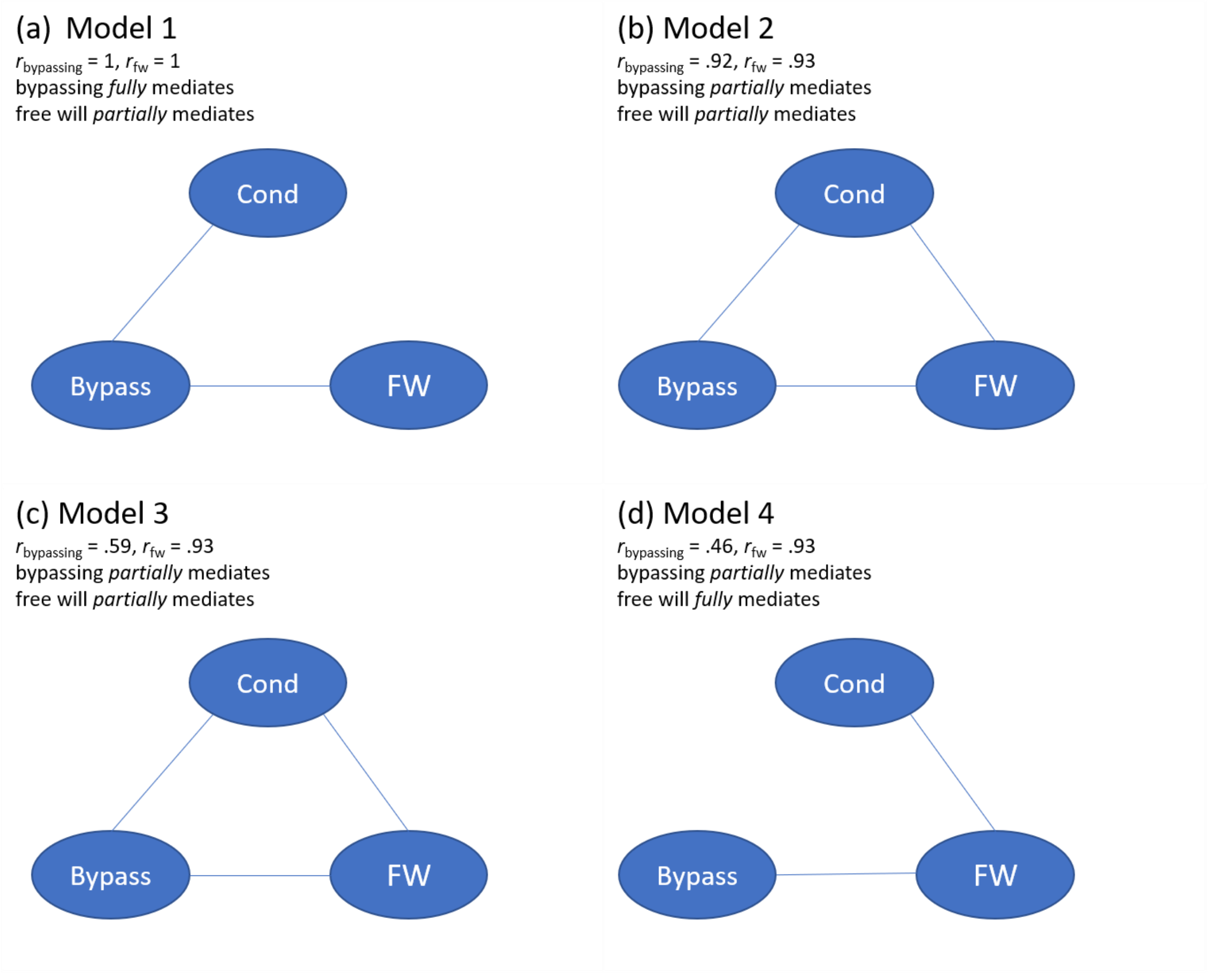
(a) Model 1
$r_{\text{bypassing}}$ = 1, r_{fw} = 1
bypassing *fully* mediates
free will *partially* mediates
Cond
Bypass
FW
(b) Model 2
$r_{\text{bypassing}}$ = .92, r_{fw} = .93
bypassing *partially* mediates
free will *partially* mediates
Cond
Bypass
FW
(c) Model 3
$r_{\text{bypassing}}$ = .59, r_{fw} = .93
bypassing *partially* mediates
free will *partially* mediates
Cond
Bypass
FW
(d) Model 4
$r_{\text{bypassing}}$ = .46, r_{fw} = .93
bypassing *partially* mediates
free will *fully* mediates
Cond
Bypass
FW

Figure 6. Causal models selected by the TETRAD Greedy Search Algorithm for simulated data, depending on the magnitude of measurement error. Measurement error is indicated by the correlation between actual data and original data (*r*).

4. An error theory for compatibilist intuitions

So far, I have argued in favor of the Bypassing Hypothesis, according to which seemingly incompatibilist answers are not genuinely incompatibilist but result from a confusion between determinism and bypassing. Thus, we could be tempted to conclude that, once that confusion is cleared, it turns out that most people have compatibilist intuitions. However, in the recent years, some have pushed the same kind of suspicions against compatibilist answers: participants' seemingly compatibilist answers might not be really compatibilists and could be explained away in a similar way we explained away incompatibilist answers. This approach is not incompatible with the Bypassing Hypothesis (it could that both seemingly compatibilist and incompatibilist answers can be explained away as mistakes), but it sheds doubt on the conclusion that laypeople are natural compatibilists.

4.1. Early error theories

Examples of such error theories for compatibilist answers are Feltz and Millan (2015)'s claim that people will ascribe "free-will-no-matter-what", or Mandelbaum and Ripley (2012)'s "Norm Broken, Agent Responsible" (NBAR) account, according to which participants will ascribe moral responsibility to agents as soon as they perceive that a norm is broken (which might explain why participants attribute moral responsibility in *concrete* cases but not in *abstract* ones). However, such accounts are not very plausible, since people do *not* automatically ascribe free will and moral responsibility, even when a norm is broken. For example, in Andow and Cova (2016), we asked participants to imagine a universe in which everyone's future is written in a magic book so that, anytime a person tries to act contrary to what is written in the book, the book will magically force this person to act against their will. Participants were presented with the story of John, whom the magical book forces to kill his wife against his will. In this case, most participants concluded that John did not act of his own free will, and was not morally responsible for killing his wife, showing that people do *not* ascribe free will "no matter what".

4.2. Folks as natural incompatibilists: the Intrusive Metaphysics Account

A more recent and subtle error theory is Nadelhoffer and colleagues (2020)'s "intrusive metaphysics" account, according to which most participants fail to accept and integrate the deterministic assumptions of thought experiments such as the *Supercomputer* case but rather “import an indeterministic metaphysics” into their interpretation of these scenarios. To put it otherwise: the apparent compatibilist judgments of participants who seem to accept that agents in deterministic universes can be free and morally responsible for their actions could be explained away by their inability to truly accept the deterministic setting of the thought-experiment and their tendency to reinterpret it in indeterministic ways (see Figure 7).

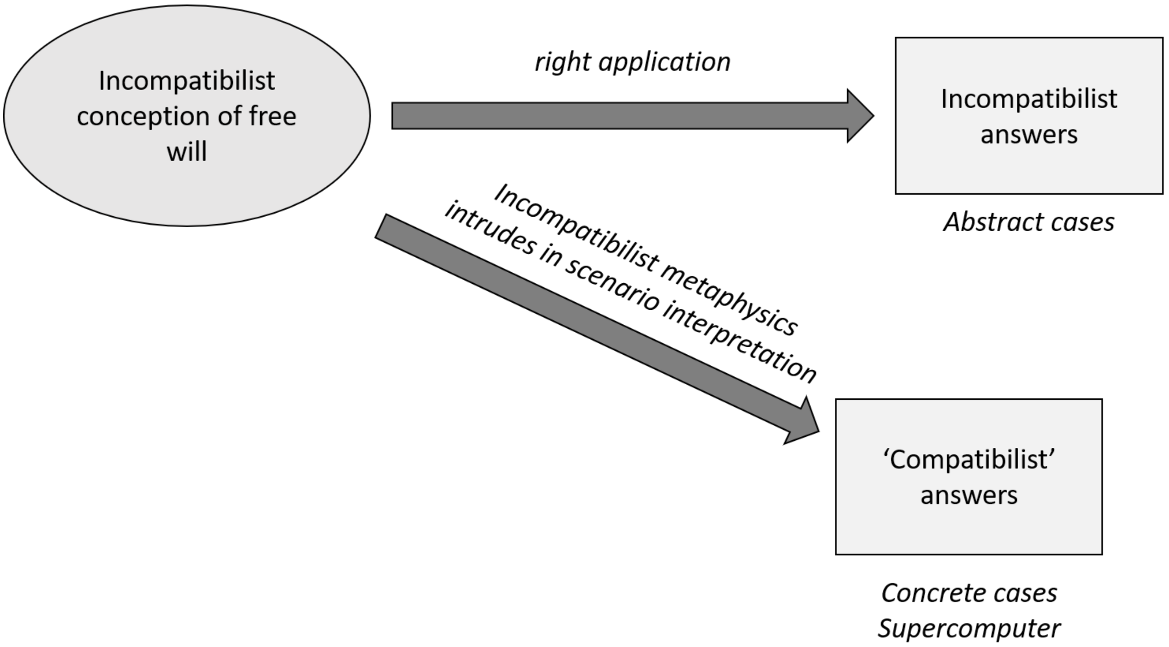

Figure 7. Nadelhoffer and colleagues’ Intrusive Metaphysics Account

To test this hypothesis, Nadelhoffer and his colleagues gave participants abstract and concrete versions of the *Supercomputer* and *Eternal Recurrence* cases developed by Nahmias and colleagues (2006). Half of participants received a version of these cases in which the agent was a robot rather than a human, but I will leave those aside to focus on participants’ answers to cases involving human agent. In these cases, participants were asked the traditional questions about the agent’s free will and moral responsibility, but also a series of questions aiming to probe to which extent they succeeded or failed to accept the deterministic setting of the scenario. For example, in the concrete version of the *Supercomputer* case, they were asked:

> **Chance:** What do you think the chances are that Jeremy will do something different than what the computer predicts he will do? (Slider scale ranging from 0 = very unlikely to 100 = very likely)

Here, I will focus on Chance as a measure of "metaphysical intrusion", as Nadelhoffer and his colleagues themselves admit that "[Chance] is perhaps our best measure of intrusion".

So what did they find? First, they found that a lot of participants indeed failed to integrate and accept the deterministic features of the vignettes. For *concrete* cases, only 47% of participants answered that the agent had *zero* chance to do something different. For *abstract* cases it was even less: 32%. Thus, it is clear that a lot of participants *do* import indeterministic assumptions in their understanding of these scenarios. This is a methodological issue future studies should keep in mind.

	Concrete cases		*Abstract cases*	
Chance	Chance = 0	Chance > 0	Chance = 0	Chance > 0
of his own free will	39% vs. 48%	14% vs. 78%	73% vs.21%	13% vs. 78%
freely	40% vs. 47%	13% vs. 77%	71% vs. 24%	14% vs. 76%
fully morally responsible for	24% vs. 62%	11% vs. 79%	44% vs. 41%	08% vs. 82%
is blameworthy / praiseworthy	31% vs. 69%	10% vs. 90%	-	-

deserves reward/punish	31% vs. 69%	15% vs. 85%	-	-
N	212	238	70	148

Table 3. Percentage of answers *below* vs. *above* the midpoint for questions about free will and moral responsibility of human agents in Nadelhoffer et al. (2020), in function of whether participants answered that the agent had zero chance to act differently. For blame/praise and and desert questions, answers were reverse-coded when the agents' action was good and I present the percentage of answers *below or equal* vs. *above* the midpoint. This re-analysis of Nadelhoffer et al. (2020)'s data was made possible by the fact that they shared their original data on osf.io/4z6r2/

But what happens if we only keep participants who answered that the agent had *zero* chance to act in another way? Do they give overwhelmingly incompatibilist answers? As we can see in Table 3, there is no simple answer to this question. For the *abstract* cases, participants' free will and moral responsibility are very low – which is not surprising, as this in line with previous results (see section 2.2). For the *concrete* cases, results are more mixed. Free will ratings (for the *free will* and *freely* questions) are split around the midpoint, with basically half of participants giving rather compatibilist answers and half of participants giving incompatibilist answers. However, when we look at questions about moral responsibility, blameworthiness/praiseworthiness and desert, roughly two thirds of participants still give compatibilist answers (which is still pretty close to the results obtained by Nahmias and colleagues). If we take into account that the number of incompatibilist answers is likely to be inflated by the fact that Nadelhoffer and colleagues did not exclude participants who confused determinism with bypassing (since they had no bypassing items), then the results are not a big deviation from what has been observed by the previous literature, as long as we focus on moral responsibility (as I have argued we should, and as I intend to do in this chapter).

Thus, a true estimate of participants' intuitions about moral responsibility would require to exclude both seemingly 'incompatibilist' answers (due to confusion with bypassing) *and* seemingly 'compatibilist' answers (due to intrusion of incompatibilist assumptions). Given the results described in Table 3, I do not think doing so will lead to results widely different from those already observed in the literature – for moral responsibility at least.

4.3. Additional limitations of the Intrusive Metaphysics Account

Additionally, the Intrusive Metaphysics Account suffers from the same problem as the Incompatibilist Model: it has very low explanatory power. As we saw in Table 3, it cannot account for the difference between abstract and concrete cases, since this asymmetry subsists even when participants who make incompatibilist assumptions are excluded. Similarly, it is not clear that it can explain the difference between neurological and psychological determinism: why would participants import indeterministic assumptions in the second but not in the first case? Here, the account needs to be refined.

Moreover, the account claims that people import indeterministic assumptions into their comprehension of scenarios because "intuitive views about the indeterministic nature of human agency influence how people understand deterministic cases like *Supercomputer*". However, this presupposes that people have an indeterministic view of human agency. But is it really the case? It is true that, when asked which of Nichols and Knobe's Universe A and B is more like ours, most people choose Universe B (the indeterministic one). However, we have seen that participants' understanding of these universes is plagued by a confusion between determinism and bypassing.

In a recent study, Giraud and Cova (in preparation; see also Giraud, 2021) asked a total of 4430 volunteers to imagine that they had been victim of a strange physics experiment and that they were now condemned to live in a temporal loop, regularly being sent back to the same point in time. Then, participants were asked the following question:

> Do you think the other persons around you in the temporal loop will always act the same way as long as you don't intervene?
>
> - Yes: they will always repeat the same behaviors
> - No: certain behaviors will sometimes be different
> - I can't say

In this case, 72.5% of participants answered 'Yes', that people in the loop will always repeat the same behaviors, which seems at odd with the idea that people have a strong indeterministic conception of human agency. Moreover, they were then asked to imagine that, in this loop, they witness two men (Bob and Charlie) repeatedly killing their respective neighbor. The sole difference was that Bob can easily be talked out of killing his neighbor, while Charlie can't. Among the participants who answered 'Yes' to the first question (and accepted determinism), 79% answered that Bob acted freely and 91% answered that he was morally responsible for killing his neighbor. Moreover, 81% answered that Charlie acted freely and 92% answered that he was morally responsible. Thus, it seems that most people do not see human agency as indeterministic, and that seeing it as deterministic does not prevent them from attributing free will and moral responsibility. As such, one of the main assumptions of the Pervasive Metaphysics account still needs to be motivated.

5. A final question in guise of conclusion

In this chapter, I tried to give an overview of current debates on laypeople's conceptions of free will and moral responsibility and their relationship with determinism. I have tried to push forward the view that people tend to be natural compatibilists, as it seems to me the best approximation of truth. However, there are still too many controversies to present this conclusion as definitive. But I would like to end on one of the main reasons why I am attracted towards natural compatibilism: I understand why people would be natural compatibilists, but I don't understand why they would be natural incompatibilists. As mentioned earlier, practices and attitudes that presuppose the attribution of moral responsibility are pervasive. They also serve an important role in regulation social relationships. As such, it is reasonable that they evolved (biologically and/or culturally) to serve certain functions. However, I can't see how these functions would be best served by imposing incompatibilist demands on moral responsibility. Indeed, from a practical point of view, such demands seem pointless. Thus, I can't see *why* and *how* we would have come to develop an intuitive incompatibilist conception of moral responsibility in the first place. To me, this is one of the most important challenges defenders of natural incompatibilism need to face.

Supplementary materials

Data for the study, simulation and re-analysis of Nadelhoffer et al. (2020)'s results presented in this paper can be found on the *Open Science Framework* at the following address: https://osf.io/8aw5e/

REFERENCES

Andow, J., & Cova, F. (2016). Why compatibilist intuitions are not mistaken: A reply to Feltz and Millan. *Philosophical Psychology*, *29*(4), 550-566.

Bergson, H. (1889/2002). *Time and Free Will: An Essay on the Immediate Data of Consciousness*. Routledge.

Berniūnas, R., Beinorius, A., Dranseika, V., Silius, V., & Rimkevičius, P. (2021). The weirdness of belief in free will. *Consciousness and Cognition*, *87*, 103054.

Björnsson, G. (2014). Incompatibilism and 'Bypassed' Agency. In Mele, A. (ed.), *Surrounding Free Will* (pp. 95–122). New York: Oxford University Press.

Björnsson, G. (2016). Outsourcing the deep self: Deep self discordance does not explain away intuitions in manipulation arguments. *Philosophical Psychology*, *29*(5), 637-653.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: a motivated account of free will belief. *Journal of Personality and Social Psychology*, 106(4), 501-513.

Cova, F. (2011). *L'architecture de la cognition morale*. Ecole des Hautes Etudes en Sciences Sociales.

Cova, F. (2014). Frankfurt-style cases User Manual: Why Frankfurt-style enabling cases do not necessitate tech support. *Ethical Theory and Moral Practice*, *17*(3), 505-521.

Cova, F. (2017). Frankfurt-style cases and the explanation condition for moral responsibility: a reply to Swenson. *Acta Analytica*, *32*(4), 427-446.

Cova, F. (forthcoming). "It was all a cruel angel's thesis from the start": Folk intuitions about Zygote cases do not support the Zygote argument. In T. Nadelhoffer (Ed.), *Advances in Experimental Philosophy of Free Will and Moral Responsibility*. London: Bloomsbury Press.

Cova, F., & Kitano, Y. (2014). Experimental philosophy and the compatibility of free will and determinism: a survey. *Annals of the Japan Association for Philosophy of Science*, *22*, 17-37.

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*(2), 851-864.

Doris, J. M., Knobe, J., & Woolfolk, R. L. (2007). Variantism about responsibility. *Philosophical Perspectives*, *21*, 183-214.

Double, R. (1996). *Metaphilosophy and free will*. Oxford University Press.

Feltz, A. (2013). Pereboom and premises: Asking the right questions in the experimental philosophy of free will. *Consciousness and Cognition*, *22*(1), 53-63.

Feltz, A. (2017). Folk Intuitions. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy). Routledge (pp.468-476)

Feltz, A., & Cova, F. (2014). Moral responsibility and free will: A meta-analysis. *Consciousness and cognition*, *30*, 234-246.

Feltz, A., & Millan, M. (2015). An error theory for compatibilist intuitions. *Philosophical Psychology*, *28*(4), 529-555.

Giraud, T. (2021). Les philosophes ne comprennent rien à la liberté. Retrieved the 25/27/2021 at https://youtu.be/FuqIY-Xf5Is

Giraud, T. & Cova, F. (in preparation). Time loops and folk intuitions about free will. Unpublished manuscript.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford University Press.

Knobe, J., & Nichols, S. (2010). Free will and the bounds of the self. In R. Kane (Ed.), *Oxford Handbook on Free Will* (2nd ed., pp. 530–554). Oxford: 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*(3), 351-368.

Mele, A. R. (2008). *Free Will and Luck*. Oxford University Press.

Miller, J. S., & Feltz, A. (2011). Frankfurt and the folk: An experimental investigation of Frankfurt-style cases. *Consciousness and Cognition*, *20*(2), 401-414.

Murray, D., & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, *88*(2), 434-467.

Nadelhoffer, T., Rose, D., Buckwalter, W., & Nichols, S. (2020). Natural compatibilism, indeterminism, and intrusive metaphysics. *Cognitive Science*, *44*(8), e12873.

Nahmias, E. (2006). Folk fears about freedom and responsibility: Determinism vs. reductionism. *Journal of Cognition and Culture*, *6*(1-2), 215-237.

Nahmias, E. & Murray, D. (2011). 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.

Nahmias, E., Coates, D. J., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest studies in Philosophy*, *31*(1), 214-242.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner 1, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*(5), 561-584.

Nahmias, E., Morris, S. G., Nadelhoffer, T., & Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research*, *73*(1), 28-53.

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin*, *40*(2), 203-216.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*(4), 663-685.

Rose, D., & Nichols, S. (2013). The lesson of bypassing. *Review of Philosophy and Psychology*, *4*(4), 599-619.

Rossi, B., & Warfield, T. A. (2017). The Relationship between Moral Responsibility and Freedom. In K. Timpe, M. Griffith & N. Levy (Eds.), *The Routledge Companion to Free Will*, pp. 612-622. Routledge.

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

Sripada, C. S. (2012). What makes a manipulated agent unfree? *Philosophy and Phenomenological Research*, *85*(3), 563-593.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies*, *173*(5), 1203-1232.

van Inwagen, P. (2008). How to think about the problem of free will. *The Journal of Ethics*, *12*(3-4), 327-341.

VanderWeele, T. J., Valeri, L., & Ogburn, E. L. (2012). The role of measurement error and misclassification in mediation analysis. *Epidemiology, 23*(4), 561.

Watson, G. (1975). Free agency. *The Journal of Philosophy*, *72*, 205–220.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*(2), 227-248.

Wolf, S. (1993). *Freedom within reason*. New York, NY: Oxford University Press.