T

Responsibility and the Relevance of Alternative Future Possibilities

Felipe De Brigard

A number of philosophers have claimed that if people are asked to consider the universe as being fully deterministic – that is, as a universe in which every event is necessarily entailed by a prior event in addition to the laws of nature – their intuitive reaction would be in line with incompatibilism about moral responsibility: i.e., they would be inclined to think that moral responsibility and determinism are incompatible[1]. However, in recent years, a number of results from experimental philosophy and psychology have cast doubt upon that claim. For example, in a series of seminal studies, Nahmias and collaborators presented participants with vignettes depicting deterministic scenarios[2]. When asked whether an agent in such scenario could have acted of her own free will, be responsible and/or deserve praise or blame for her actions, the majority of participants answered affirmatively. These results led Nahmias and colleagues to suggest that contrary to the received, a priori view among philosophers, people may actually be compatibilists.

Soon after, a number of studies challenged this conclusion. First, Nichols and Knobe reported results from a series of studies in which participants were presented with vignettes depicting fully deterministic scenarios[3]. How-

[1] R. Kane, *The Significance of Free Will*, Oxford University Press, Oxford-New York 1996; D. Pereboom, *Living without Free Will*, Cambridge University Press, Cambridge 2001.

[2] E.A. Nahmias-S.G. Morris-T. Nadelhoffer-J. Turner, *Surveying freedom: Folk intuitions about free will and moral responsibility*, in «Philosophical Psychology», 18 (2005), n. 5, pp. 561-584; Idd., *Is incompatibilism intuitive?*, in «Philosophy and Phenomenological Research», 73 (2006), n. 1, pp. 28-53.

[3] S. Nichols-J. Knobe, *Moral responsibility and determinism: The cognitive science of folk intuitions*, in «Nous», 41 (2007), n. 4, pp. 663-685.

ever, half of the participants received vignettes couched in abstract and emotionally neutral terms whereas the other half received vignettes couched in concrete and emotionally salient terms. They found that participants who read the deterministic scenarios described in concrete and emotionally salient terms were more likely to align their judgments of responsibility with compatibilism. In contrast, participants who read the deterministic scenarios described in abstract and emotionally neutral terms, made judgments that aligned with incompatibilism. Importantly, a more recent study suggests that this effect is evident across many cultures[4]. A second series of experiments conducted by Roskies and Nichols also challenged Nahmias et al.'s claim that people are naturally compatibilists[5]. In their study, Roskies and Nichols asked participants to read a vignette, similar to those employed in the previous studies, depicting a fully deterministic scenario. However, half of the participants were asked to imagine the described event occurring in a possible but non-actual world while the other half were told that the described event occurs in the actual world. Their results suggest that participants are more likely to give incompatibilist answers when they read vignettes depicting deterministic scenarios in a possible yet non-actual world whereas in scenarios described as occurring in the actual world their responses align with compatibilism. Finally, results from a study by Nahmias, Coates and Kvaran suggest that when presented with deterministic scenarios described in purely reductionistic terms, participants' judgments of responsibility align with compatibilism if the terms on the vignette are concrete and emotionally salient, but this is not the case if the vignettes are abstract and emotionally neutral[6].

A number of proposals have tried to accommodate these conflicting results. According to one proposal[7], the results of these studies could be accounted for if we assume a more basic psychological distinction between two distinct cognitive systems underlying our judgments of moral responsibility. On the one hand, there is a concrete system in charge of generating

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

[5] A. Roskies-S. Nichols, *Bringing responsibility down to earth*, in «Journal of Philosophy», 105 (2008), n. 7, pp. 371-388.

[6] E. Nahmias-D.J. Coates-T. Kvaran, *Free will, moral responsibility, and mechanism: Experiments on folk intuitions*, in «Midwest Studies in Philosophy», 31 (2007), pp. 214-242. See also F. De Brigard-E. Mandelbaum-D. Ripley, *Responsibility and the brain sciences*, in «Ethical Theory and Moral Practice», 12 (2009), n. 5, pp. 511-524.

[7] S. Nichols-J. Knobe, *op. cit.*

judgments of moral responsibility when facing reductionistic, mechanistic, concrete and emotionally loaded deterministic scenarios. On the other hand, there is an abstract system in charge of producing judgments of responsibility for non-reductionistic, non-mechanistic, abstract and emotionally neutral deterministic scenarios. Indeed, Sinnott-Armstrong (2008)[8] has suggested that these two systems may be underwritten by the widely accepted distinction between episodic and semantic memory systems. More recently, a different proposal has been put forth by Murray and Nahmias[9]. According to their view, people are naturally compatibilists; their apparent incompatibilist judgments occur as a result of participants misinterpreting determinism as implying that the agent's mental states are bypassed in the causal chain leading up to the action. Thus, participants' incompatibilist intuitions can be explained away as an error in judgment. Needless to say, the debate as to whether peoples' judgments of responsibility align with compatibilism or incompatibilism in deterministic scenarios is far from being settled[10].

However, results from a recent study by De Brigard and Brady may pose an unexpected problem to the ecological validity of many of the studies reported in this debate[11]. In agreement with the way philosophers talk about the problem of free will, determinism and responsibility, experimental psychologists and philosophers have focused their efforts in exploring peoples' judgments of responsibility in scenarios where the only information that is provided pertains to the causal history preceding the agent's action. Specifically, researchers have been interested in determining which sorts of considerations about actual (or counterfactual) *past* events that bring about the agent's action influence peoples' judgments of responsibility in deterministic scenarios. But the fact that traditionally philosophers have only cared about the events that precede the agent's action does not mean that ordinary folk make the same assumption. There are a number of philosophical, moral and legal reasons to dismiss the import of

[8] W. Sinnott-Armstrong, *Abstrac+ Concrete = Paradox*, in J. Knobe-S. Nichols (eds.), *Experimental philosophy*, Oxford University Press, New York 2008, pp. 209-230.

[9] D. Murray-E. Nahmias, *Explaining away incompatibilist intuitions*, in «Philosophy and Phenomenological Research», 88 (2014), n. 2, pp. 434-467.

[10] S. Nichols, *Experimental philosophy and the problem of free will*, in «Science», 331 (2012), pp. 1401-1403; E. Nahmias, *Free Will and Moral Responsibility*, in «Wiley Interdisciplinary Reviews: Cognitive Science», 3 (2012), pp. 439-449.

[11] F. De Brigard-W. Brady, *The effect of what we think may happen on our judgments of responsibility*, in «Review of Philosophy and Psychology», 4 (2013), n. 2, pp. 259-269.

the consequences that may ensue if a person is held responsible at the present time. But there is no a priori reason to believe that ordinary people share those reasons and that they do not consider possible future events when judging if a person is or not responsible – even under fully deterministic and emotionally salient scenarios. Whether or not the folk's judgments about responsibility in fully deterministic scenarios are influenced by considerations about possible future events that may ensue as a result of holding an agent responsible is an open empirical question.

This issue is precisely what De Brigard and Brady set up to explore. In three between-group experiments they presented participants with mechanistic, reductionistic, emotionally loaded, and concrete deterministic scenarios of the sort that, consistently, have led participants to generate judgments of responsibility in line with compatibilism[12]. However, they manipulated whether possible future consequences that may ensue as a result of holding the agent responsible either improve or worsen the situation of an innocent third-party. Here, for instance, is the vignette read by the participants in the first experiment:

> Mary is the single mother of two: Mark, 7, Sally, 4. Mary works most of the day, and although she is known for being fairly patient and good natured, over the last year she has exhibited some unusually aggressive behavior toward her neighbor. Last week, when she came back from work late at night, she couldn't drive into her garage because her neighbor had blocked her driveway with his new BMW. Enraged, she stepped on the gas pedal and crashed her car into her neighbor's. Unfortunately, her neighbor was still inside the car (it was too dark for anyone to see him), and both his legs were seriously broken in several places. Now he is not only suing her for several thousand dollars, but he's also pressing charges. However, a neurologist examined her brain and discovered that, in the last year, Mary has been developing a rare tumor in her frontal lobe. Since the frontal lobe is necessary for emotional suppression – that is, the capacity to control one's emotions – the neurologist claims that, unlike a healthy person, Mary was completely unable to control her rage and her desire to smash the car. "In fact", he says, "any person with this kind of tumor", facing the exact same situation, would have done exactly what Mary did. She couldn't have done otherwise. "If Mary is found responsible for her actions, she may be sent to a federal medical facility for the next 6 months". There she could receive medical treatment, but she won't be able to see her children[13].

[12] *Ibidem*.
[13] *Ivi*, p. 262.

Half of the participants were randomly assigned to the *Better* condition, in which the vignette concluded with the following sentence:

> Fortunately, during that time, they would be living with Aunt Elizabeth, in what might be a much better environment for them.

The other half were assigned to the *Worse* condition, in which the vignette concluded with the following sentence:

> Unfortunately, during that time, they would be living with Social Services, in what might be a much worse environment for them.

Immediately after participants were asked to rate, on a 1-7 Lickert scale, whether or not they agreed or disagreed with the statement "Mary is morally responsible for crashing her car into her neighbor's". The results indicate that participants were significantly more likely to say that Mary was responsible in the *Better* condition (M = 5.30, SD = 1.2) than in the *Worse* condition (M = 3.15, SD = 1.7). These results suggest that, even under fully deterministic and emotionally-salient scenarios, when participants considered that the situation of an innocent third-party may worsen as a result of holding an agent responsible at the present time, their judgments are more aligned with compatibilism. However, when they considered that the condition of an innocent third-party may improve as a result of holding the agent responsible, their judgments were more in line with incompatibilism.

Since studies employing vignettes involving neural pathologies have produced conflicting results[14], De Brigard and Brady conducted two follow-up experiments in which the agent did not have a neural pathology[15]. In the first follow up, which was also a between subjects experiment, participants read a vignette similar to the one employed in the first experiment, except that this time the concrete and deterministic character of the description of the events leading up to the action was captured by assuming that Mary was wearing a brain monitoring system that recorded her brain activity. A neuroscientist then interpreted the data recorded from Mary's brain activity and concluded that the brain events leading up to Mary's action were completely determined and that she could not have done otherwise. As before, half of the participants were assigned to the *Better* condition and the other half were assigned to the *Worse* condition.

[14] F. De Brigard-E. Mandelbaum-D. Ripley, *op. cit.*
[15] F. De Brigard-W. Brady, *op. cit.*

The results of this second experiment revealed that participants were more likely to say that Mary was responsible for crashing her car into the neighbor's in the *Better* condition (M = 5.75, SD = 1.26) than in the *Worse* condition (M = 4.38, SD = 1.76) This suggests that participants were more prone to hold an agent responsibility if they considered that an innocent third-party may possibly be better off in the future as a result. Conversely, if the innocent third-party could end up worse off, participants' judgments of responsibility did not differ from the midpoint, suggesting that albeit not enough to exculpate the agent, considering this undesirable possible future consequences was sufficient to prevent participants from generating full-fledged compatibilist (or incompatibilist) judgments.

Finally, to explore whether or not the effect of considering possible consequences for innocent third-parties is a more pervasive characteristic of our judgments of responsibility, De Brigard and Brady conduced one final experiment in which the narrative about determinism was removed[16]. As before, half of the participants received a *Better* vignette, while the other half received a *Worse* vignette. Consistent with the results from their second experiment, the results of this final experiment revealed that participants were more likely to attribute responsibility to Mary if her children could be better off as a result of she going to a correctional facility (M = 6.17; SD = 1.24) than if they may be worse off (M = 4.46; SD = 2.21). Thus, taken together, the results of these three experiments strongly suggest that when assessing whether an agent is or not responsible for a particular action, people may consider possible future consequences for innocent third-parties that may be brought about as a result of holding the agent responsible at a present time. Moreover, this effect appears to be independent of whether or not the description of the conditions under which the agent acts is fully deterministic, mechanistic, reductionistic, and emotionally laden.

What may account for these results? The proposal I would like to put forth builds upon a recent and provocative paper by Phillips, Luguri and Knobe[17]. Their paper deals with the well-known phenomenon that moral judgments seem to influence non-moral assessments in a variety of domains. For instance, in a pioneer study, Knobe demonstrated that participants were more likely to say that an agent brought about a side effect he didn't care about when said side effect was morally wrong but not when it

[16] *Ibidem*.

[17] J. Phillips-J. Luguri-J. Knobe, *Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities*, in «Cognition», 145 (2015), pp. 30-42.

was morally right[18]. Relatedly, Phillips and Knobe conducted a study in which participants read a vignette depicting a scenario in which the captain of a ship saves its vessel from sinking by throwing his wife's cargo overboard (the morally neutral condition) or by throwing his wife overboard (the morally bad condition)[19]. Overall, participants were more likely to say that the captain was forced to throw something overboard in the morally neutral condition than in the morally bad condition. To explain these – and other related – results, Phillips and collaborators suggest, and offer evidence in favor of, the claim that moral considerations influence the kinds of possibilities people consider relevant when generating judgments about different notions across a number of distinct domains, such as intentional action, force, causation and doing/allowing[20]. More specifically, their suggestion is that «people show a general tendency to regard alternative possibilities as more relevant to the extent that they involve replacing morally bad things in the actual world with morally good alternatives»[21].

A similar explanation may be available for the effects uncovered by De Brigard and Brady[22]. Their results suggest that if a morally bad consequence could be brought about in the future as a result of holding an agent responsible at a present time, then participants are less likely to hold the agent responsible than if a morally good consequence were to be brought about. In agreement with Phillips and colleagues' proposal, one can hypothesize that this effect is due to a shift on attention toward relevant future possibilities that may be considered by the participants[23]. Thus, in the *Worse* condition, the morally bad effect on Mary's children renders certain possible future consequences more relevant, such as them having to live with someone they do not know, getting behind in school, or perhaps being mistreated in Social Services. Possible good consequences that may follow from this bad effect on Mary's children are not rendered relevant, thus they are not considered plausible. Because these bad consequences are ren-

[18] J. Knobe, *Intentional action and side effects in ordinary language*, in «Analysis», 63 (2003), n. 3, pp. 190-194.

[19] J. Phillips-J. Knobe, *Moral Judgments and Intuitions about Freedom*, in «Psychological Inquiry», 20 (2009), pp. 30-36. See also L. Young-J. Phillips, *The Paradox of Moral Focus*, in «Cognition», 119 (2011), pp. 166-178.

[20] J. Phillips-J. Luguri-J. Knobe, *op. cit*. See also D. Pettit-J. Knobe, *The Pervasive Impact of Moral Judgment*, in «Mind & Language», 24 (2009), n. 5, pp. 586-604.

[21] J. Phillips-J. Luguri-J. Knobe, *op. cit.*, p. 40.

[22] F. De Brigard-W. Brady, *op. cit.*

[23] J. Phillips-J. Luguri-J. Knobe, *op. cit.*

dered more plausible in the *Worse* conditions, participants may be motivated to prevent them from happening by way of judging the responsibility of the subject less harshly. Conversely, in the *Better* condition, the morally good effect on Mary's children renders other good consequences as being more relevant, like the fact that the nice aunt Elizabeth may provide a nurturing home for them, and would probably prevent them from getting in trouble or behind in school. Because good consequences are now rendered relevant – thus likely – people may be less inclined to mitigate Mary's responsibility – as there is less of an urge to prevent this outcome to occur.

Needless to say, this is merely a hypothesis. While it is inspired by Phillips and colleagues' recent proposal[24], it differs from theirs in an important respect. In their proposal, moral judgments influence the kinds of *counter-factual thoughts* participants entertain when assessing a certain situation. In the current interpretation of De Brigard and Brady's results[25], moral judgments influence *pre-factual thoughts* participants entertain when assessing Mary's moral responsibility. In other words, while their proposal states that moral judgments increase the relevance of certain thoughts about alternative ways past events could have occurred, the current proposal suggest that they can also render as relevant certain thoughts about how possible future events may unfold. Although it is so far an untested hypothesis, some extant evidence suggest that it may be promising, as it turns out that there is much in common between the neural and cognitive mechanisms underlying our capacity to entertain episodic future and counterfactual thoughts[26]. As such, the temporal dimension of the hypothetical simulation participants engage in during their judgments may not be critical[27]; what matters is the degree to which the moral character of the initially suggested possibility renders other possibilities as more or less relevant or plausible.

This need not be the whole explanation, of course. Extant evidence also suggests that our impulse to blame the perpetrator influences our attribu-

[24] *Ibidem*.

[25] F. De Brigard-W. Brady, *op. cit.*

[26] F. De Brigard-D. Addis-J.H. Ford-D.L. Schacter-K.S. Giovanello, *Remembering what could have happened: Neural correlates of episodic counterfactual thinking*, in «Neuropsychologia», 51 (2013), n. 12, pp. 2401-2414; D.L. Schacter-R. Benoit-F. De Brigard-K.K. Szpunar, *Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions*, in «Neurobiology of Learning and Memory», 117 (2015), pp. 14-21.

[27] F. De Brigard-B.S. Gessell, *Time is not of the essence: Understanding the neural correlates of mental time travel*, in S.B. Klein-K. Michaelian-K.K. Szpunar (eds.), *Seeing the Future: Theoretical Perspectives on Future-Oriented Mental Time Travel*, Oxford University Press, Oxford-New York 2016.

tions of free will and responsibility[28]. Notice, however, that this account does not conflict with the proposal put forth here, as each suggests a different process influencing our judgments of free will and responsibility. On the account put forth here, the main process is attention to relevant possibilities, whereas in the impulse-to-blame account the main process appears to be emotional. Clearly, further research is needed to fully understand the interaction between the impulse to blame and the relevance of alternative possibilities as factors influencing people's judgments of free will and responsibility.

Finally, in addition to offering a possible explanation of De Brigard and Brady's findings, it is worth mentioning at least two important methodological consequences that follow from them for both experimental philosophy and psychology of free-will and determinism[29]. First, both experimental philosophers and psychologists may want to take note of the relevance of possible future consequences when asking participants to assess the degree of responsibility of an agent in particular deterministic scenarios. The history of philosophy is full of prescriptive reasons as to why such consequences should not be taken into consideration when judging whether or not an agent is responsible for an action. However, such prescriptive considerations need not be entrenched in the psychological processes ordinary folk engage in when judging whether or not an agent is responsible. After all, our concept of responsibility presumably developed to play a social role – perhaps to curb people's behavior after a condemnable action, or to draw attention to the untrustworthiness of the agent, or who knows. But either way, it would be a mistake to assume that considerations about possible future events that we, philosophers or legal theorists, have learned to disregard on the basis of some prescriptive reason are also disregarded as a matter of course by ordinary people when judging the responsibility of an agent.

The second consequence follows, by way of generalization, from the first one: when designing vignettes to test people's intuitions about one or another notion – such as determinism, responsibility, free-will, and so forth – it is important not to mistakenly assume that our philosophical reasons for thinking that certain details are not relevant for the vignette are

[28] M.D. Alicke, *Culpable control and the psychology of blame*, in «Psychological Bulletin», 126 (2000), pp. 556-574; C.J. Clark-J.B. Luguri-P.H. Ditto-J. Knobe-A.F. Shariff-R.F. Baumeister, *Free to punish: A motivated account of free will belief*, in «Journal of Personality and Social Psychology», 106 (2014), pp. 501-513.

[29] F. De Brigard-W. Brady, *op. cit.*

also psychological reasons for thinking so. After all, one of the major downfalls of conducting research with these sorts of vignettes is that the researcher has only indirect control of the independent variable: she can manipulate what participants read, not what they think, and often what they think involves more than what they read. In experimental settings researchers work hard to keep background conditions as stable as possible in order to increase the probability that the intervention on the independent variable is predictive of the change in the dependent variable. The effects revealed by De Brigard and Brady suggest that something that was considered stable and irrelevant for the manipulation – i.e., considerations about possible future events – may actually have an effect on the dependent variable. As such, this finding constitutes an avenue for future research but also a possible worry about prior effects[30].

To conclude, let me summarize what I attempted to do in the current paper. I started off by briefly reviewing a number of recent results from experimental philosophy and psychology suggesting that, under certain conditions, people's intuitive compatibilist judgments shift toward incompatibilism even when considering fully deterministic scenarios. To account for these results, a couple of proposals have been put forth, including the suggestion that the kinds of cognitive processes involved in thinking about concrete, reductionistic, mechanistic and emotionally-laden deterministic scenarios are different from the kinds of cognitive processes involved in thinking about abstract, non-reductionistic, non-mechanistic and emotionally-neutral scenarios. However, recent findings from De Brigard and Brady put pressure on this proposal, as alternative future possibilities seem to affect participant's judgments of responsibility from compatibilist to incompatibilist even when they are presented with concrete, reductionistic, mechanistic, and emotionally-laden deterministic scenarios. As a result, building upon a recent proposal by Phillips and colleagues, a different account was put forth: that bringing attention to either morally bad or morally good outcomes renders certain related possibilities as more or less likely, thus as more or less relevant for considering whether or not the agent is responsible. Finally, I drew a couple of methodological sugges-

[30] *Ibidem*. It is worth noting that others have expressed skepticism as to whether responses to moral dilemmas in experimental settings actually reflect responses to similar situations in real-life settings (cf. G. Kahane-J.A.C. Everett-B.D. Earp-M. Farias-J. Savulescu, *"Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good*, in «Cognition», 124 (2015), pp. 193-209).

tions from these results, in the hopes that bringing attention to potential confounds in extant experimental designs can help to motivate more ecologically valid studies moving forward[31].

Abstract

In the past decade, philosophical and psychological research on people's beliefs about free will and responsibility has skyrocketed. For the most part, these vignette-based studies have exclusively focused on participants' judgments of the causal history of the events leading up to an agent's action and considerations about what the agent could have done differently in the past. However, recent evidence suggests that, when judging whether or not an individual is responsible for a certain action – even in concrete, emotionally laden and fully deterministic scenarios – considerations about alternative future possibilities may become relevant. This paper reviews this evidence and suggests a way of interpreting the nature of these effects as well as some consequences for experimental philosophy and psychology of free will and responsibility going forward.

Keywords: experimental philosophy; experimental psychology; free will; responsibility.

Felipe De Brigard
Department of Philosophy
Duke University, Durham, North Carolina
felipe.debrigard@duke.edu

[31] Thanks to Paul Henne and two anonymous reviewers for helpful comments.