“It was all a cruel angel’s thesis from the start”: Folk intuitions about Zygote cases do not support the Zygote argument

Florian Cova

Penultimate draft, final version to appear in Nadelhoffer, T. & Monroe, A. (Ed.), *Advances in Experimental Philosophy of Free Will and Responsibility*. Bloomsbury.

1. Introduction: Free Will and Manipulation Arguments

For the purpose of the present paper, let’s define free will as the kind of control one needs to exert upon one’s actions to be morally responsible for them (Mele, 2006). One central question regarding free will is whether free will is compatible with determinism, where determinism should be understood in the following way: a system (for example, the world) is deterministic when it obeys rules such that the state of this system at any moment *t* could be in principle deduced from the conjunction of these rules and a description of the state of this system at any anterior moment *t-n*. To put it otherwise: a system is deterministic when, given the past and the rules that govern this system, anything that happens in this system *has to* happen the way it does, and not in another way.

Determinism is thus the thesis that the world is such a system (at least at the level relevant to human action), and thus that human actions are *determined*: for every human action, it would be in principle possible (if one had perfect knowledge of the laws of nature and prior states of the world) to understand why one had to perform *this* action and not *another one*. For a certain number of philosophers, called *incompatibilists,* a deterministic world is incompatible with free will: if it is the case that, given the past, each agent had to act the way they did and it was impossible for them to act otherwise, then these agents cannot be morally responsible for their action. They are opposed by *compatibilists*. Compatibilist philosophers think that determinism is no threat to free will and that both can perfectly coexist.

But why think free will and determinism are incompatible (or compatible)? Philosophers on both sides have developed a wide array of arguments for and against the compatibility of free will with moral responsibility. In this paper, we will focus on one particular class of arguments for the *incompatibility* of free will and determinism: *Manipulation Arguments* (Mickelson, 2017; Mele, 2019).

Manipulation arguments can be described in the following way. They are arguments that start from the (rather uncontroversial) claim that (some class of) manipulated agents are neither free nor morally responsible for their actions. Then, they move from this claim to the claim that agents in a deterministic world are neither free nor morally responsible for their actions (a famous example is Pereboom’s four-case argument; see Pereboom, 1995). This move can be operated in two different ways.

A first class of manipulation arguments (that we can call *direct* or *transfer* manipulation arguments) operates in the following way: after obtaining our agreement that agents in manipulation cases are neither free nor morally responsible for their actions, they point to the fact that there is no *relevant* difference between manipulation cases and the cases of agents living in deterministic universes. Thus, we should conclude that, since there is no relevant difference between the two cases, and it is clear that agents in manipulation cases are not morally responsible for their actions, then agents in deterministic universes are not morally responsible for their actions. The argument leads us to *directly transfer* our intuitions about manipulation cases to agents living in a deterministic world.

However, several philosophers have been pointing out a serious limitation of such *direct* manipulation arguments: even if they were successful, they would only teach us that agents in deterministic universes are not morally responsible for their actions, but they would not tell us *why*. More precisely, showing that agents in deterministic are not morally responsible is not the same as showing that *determinism itself* is an obstacle to moral responsibility: it could be that deterministic universes also have some other properties and that these other properties are the one making moral responsibility impossible, rather than determinism. Thus, certain philosophers have argued that manipulation arguments do not show that determinism precludes moral responsibility, but that moral responsibility is impossible *tout court* - which entails that agents in deterministic worlds are not morally responsible for their actions, but not for the reasons incompatibilists claimed (Levy, 2011; Mickelson, 2015).

These shortcomings can be avoided by taking another approach to manipulation arguments - an *indirect* or *explanatory* one.[1] This approach consists in proposing an explanation of why agents in manipulation cases are not morally responsible for their actions, then showing that this explanation also applies to cases of agents living in a deterministic universe (see Mickelson, 2016). Of course, for the manipulation argument to vindicate incompatibilism, the explanation must point at a certain feature that is implied by determinism, if not determinism itself.

One way to counter such *indirect* or *explanatory* manipulation argument is to show that our intuition that manipulated agents are not morally responsible for their action has nothing to do with determinism or something implied by determinism. For example, Mele (2005) rejects Pereboom's Four Cases arguments by showing that introducing determinism in Pereboom's manipulation cases does not change our intuitions about manipulated agents' moral responsibility, which suggests that the right explanation for our verdict has nothing to do with determinism.

Overall, potential responses to manipulation arguments can be classified in two main categories (Kane, 1985). *Hard-line responses* consist in denying that agents in manipulation cases are not morally responsible for their actions. *Soft-line responses* consist in accepting the premise that agents in manipulation cases are not morally responsible for their actions but denying that this

[1] Mickelson (2016) distinguishes a third type of manipulation argument, based on *generalization*. For the sake of simplicity, and because I won't need them for my discussion of experimental results, I won't develop such arguments here.

verdict can be transferred to deterministic cases (either because there are relevant differences between both cases or because the explanation does not apply to deterministic cases).

2. Experimental philosophy and manipulation arguments

In the past years, experimental philosophers have collected data on people's intuitions about manipulation cases and have used it to push forward both *hard-line* and *soft-line* responses.

2.1. Experimental philosophy and the hard-line response

Feltz (2013) investigated folk intuitions about Pereboom's four cases argument. After controlling for a potential confounding factor that had nothing to with determinism (namely, the presence of the manipulator's intention), Feltz observed that participants tended to consider that agents in manipulation arguments were in fact morally responsible for their actions. Feltz thus argued that his results might justify a hard-line response to Pereboom's argument: if we do not have the intuition that agents in manipulation cases are not morally responsible for their actions, then the very first premise of the argument is threatened.

2.2. Experimental philosophy and the soft-line retort

Other experimental studies have provided the ground for soft-line responses to manipulation arguments. In a seminal paper, Sripada (2012) argued that, to the extent that participants consider agents in manipulation case not to be morally responsible for their actions, it is because they think that manipulated agents differ from everyday agents in one crucial respect: they do not act in accordance with their "deep self" (i.e. their 'real', deeply-held values and attitudes). Sripada conducted a first study in which participants receive either the *Manipulation* or *Control* case (see Table 1).

Manipulation	*Control*
One day, Bill sees a woman named Mrs. White as she is jogging in the park. Bill hates this woman, and deliberates about what to do. After weighing his options, Bill decides he should kill her. Bill's mind is not clouded by rage or other extreme emotions. Rather, Bill thinks clearly and carefully about his own desires and values, and only then makes a decision. After he kills Mrs. White, Bill reflects on his action. He wholeheartedly endorses what he has done. But there is more you need to know about Bill, and how he came to be the person that he is now: There is a man named Dr. Z who is a scientific genius and who is an expert at indoctrination. Dr. Z hates Mrs. White and formed the following plan. Dr. Z would take an infant from an orphanage and raise the child himself. He would teach and reward just the right behaviors in the	

child so the child would hate Mrs. White and want her dead. He would script all the major events in the child's life to nurture and cultivate in the child the goal of doing whatever it would take to kill Mrs. White. Dr. Z tried this plan previously on five other children, and each time the child grew up to kill Dr. Z's intended targets.	
Dr. Z implemented his plan for Bill. He took Bill from an orphanage when Bill was an infant. The plan worked—once Bill had grown up, Bill had the desire to do whatever it takes to kill Mrs. White. Dr. Z's plan was kept completely hidden from Bill. Bill never knew that Dr. Z implemented the plan.	Dr. Z implemented his plan for Bill. He took Bill from an orphanage when Bill was an infant. The plan worked—once Bill had grown up, Bill had the desire to do whatever it takes to kill Mrs. White. Dr. Z's plan was kept completely hidden from Bill. Bill never knew that Dr. Z implemented the plan.

Table 1. Vignettes used by Sripada (2012).

In both cases, participants were asked about the agent's moral responsibility, but also about the concordance between his actions and his Deep Self (e.g. "Bill's killing of Mrs. White does not reflect the kind of person who he truly is deep down inside"). The results showed that participants considered Bill in the *Manipulation* case to be less responsible for his actions than Bill in the *Control* case, but that this difference was mediated by participants' deep self ratings: participants considered that Bill in the *Manipulation* case was not necessarily acting in accordance with his Deep Self.

In a second study, Sripada introduced a *Modified* manipulation case, which was identical to the *Manipulation* case, except that the following paragraph was added at the end:

> Bill is like anyone else in many respects. As he was growing up, Bill was educated about morality, the difference between right and wrong, and various ways he might conduct his life. Additionally, Bill was not simply fed lies about Mrs. White—he knows the truth about who she is and he knows exactly why he dislikes her. Bill is not a robot who simply does as others instruct. Nor is he under the grip of an irresistible impulse. Rather, Bill is a person, with desires, values, hopes, and dreams just like anyone else. But Bill's desires include killing Mrs. White. And his core values permit killing Mrs. White. So that is exactly what he does.

Compared to the first *Manipulation* case, adding this information about Bill's Deep Self led participants to see Bill as more responsible for his action. In fact, a majority of participants now answered that Bill killed Mrs. White *of his own free will*.

Together, these results suggest the following soft-line responses to manipulation arguments: to the extent that participants see manipulated agents as unfree, it is not because they perceive them as determined, but as acting against their own deeply-held values. Thus, unless someone is able to show that determinism precludes people from acting on the basis of their deeply-held values, we are not justified in transferring our intuitions about manipulation cases to cases involving agents in deterministic worlds.

Of course, one might object that these results are philosophically relevant. After all, we should take as starting point the intuitions of experts, and not of untutored, naïve laypeople. However, it is not clear that this response is as effective as it might seem at first sight: past research in the philosophy of free will has tried to keep as close as possible to our intuitive conception of moral responsibility. Indeed, proving that moral responsibility is or is not possible might not prove practically relevant and worthwhile if the conception of moral responsibility one is using has cut all relationships to this intuitive conception. After all, if people's intuitions about manipulation case are that manipulated agents are free and morally responsible, why should they even begin tow worry about Pereboom's argument?

But even if consider that only expert intuitions are relevant to philosophical discussions about free will, this does not undermine Sripada's argument. Indeed, Sripada's argument is not that people's intuitions about Manipulation cases are not in line with those of experts', but that intuitions about Manipulation cases (when they agree with those of experts') are not explained by determinism but by considerations compatibilism can accommodate. Of course, one could still object that experts' intuitions are driven by completely different mechanisms than those of laypeople. However, this seems a gratuitous assertion for which there is at the moment no empirical evidence – and thus, some dose of skepticism is warranted. Moreover, we will see in the final discussion that some of the results presented in this paper warrants the idea that laypeople and experts' intuitions about free will seem to follow similar patterns.

3. Here comes a new challenger: the Zygote Argument

So far, all cases of manipulation that experimental philosophers have put to the test share one common feature: agents are manipulated *after their birth*. This might seem like a trivial observation, but one might actually argue that this is what leads participants to see a dissonance between manipulated agents and their Deep Self: after all, they already had desires and values before being manipulated. Thus, participants might see manipulated agents as divided or torn between two sets of desires and values: their original ones and the one that have been added by manipulation. This conception of agents' minds as divided might then drive them to consider the new desires and values added by manipulation as non-authentic and to some extent akin to brainwashing (Sripada, 2012).

If this hypothesis is true, then this means that it would be possible to avoid the kind of objection Sripada and other experimental philosophers have raised to manipulation arguments by turning to manipulation arguments in which agents are manipulated *before* being born, and thus before having any kind of preexisting desires and values that might be considered as the agents' 'true' desires and values. One such case of argument is Alfred Mele's Zygote argument.

The Zygote argument begins with a thought experiment:

> Consider the following story. Diana creates a zygote *Z* in Mary. She combines *Z*'s atoms as she does because she wants a certain event *E* to occur thirty years later. From her knowledge of the state of the universe just prior to her creating *Z* and the laws of nature of her deterministic universe, she deduces that a zygote with precisely *Z*'s constitution located in Mary will develop into an ideally self-controlled agent who, in thirty years, will judge, on the basis of rational deliberation, that it is best to *A* and will *A* on the basis of that judgment, thereby bringing about *E*. If this agent,

> Ernie, has any unsheddable values at the time, they play no role in motivating his *A*-ing. Thirty years later, Ernie is a mentally healthy, ideally self-controlled person who regularly exercises his powers of self-control and has no relevant compelled or coercively produced attitudes. Furthermore, his beliefs are conducive to informed deliberation about all matters that concern him, and he is a reliable deliberator. So he satisfies a version of my proposed compatibilist sufficient conditions for having freely *A*-ed. (Mele, 2008:279).

For the Zygote argument to begin, we have to assume that you have the intuition (or that you agree) that Ernie is not morally responsible for his A-ing in this case. Once this crucial premise accepted, the Zygote argument runs like this:

1) Because of the way his zygote was produced in his deterministic universe, Ernie is not a free agent and is not morally responsible for anything.
2) Concerning free action and moral responsibility of the beings into whom the zygotes develop, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe.
3) So determinism precludes free action and moral responsibility. (Mele, 2008:280)

Thus, from the intuition that agents in Zygote cases (such as Ernie's) are not morally responsible for their actions, the Zygote argument concludes that moral responsibility is incompatible with determinism.

4. Goal of the present paper and some methodological considerations

In this paper, my goal is to empirically investigate people's intuition about Zygote cases. My main driving hypothesis is that, even if manipulation intervened before the agent's birth, people *still have* (at least implicitly) the intuition that manipulated agents in Zygote cases are not able to act on the basis on their true desires and values, and that this explains why some people have the intuition that agents in Zygote cases are not morally responsible for their actions.

Before moving on to the empirical part of the paper and to my data-driven assessment of Mele's Zygote argument, there are a few methodological points I would like to highlight, even if I do not have enough space to fully develop them:

1) I will not focus on participants' judgments about free will, but on their judgments about moral responsibility. The reason why is because I am primarily interested in FREE WILL, but FREE WILL in the sense it is currently discussed by most philosophers, that is: the kind of control one must exert upon one's actions to be morally responsible for them. There might be a folk concept of free will (though it is not clear that there is one in every language; see Berniūnas et al., 2021), but it is not directly superposable to the one I am interested in. For example, in Nahmias and colleagues (2006)'s experiments, free will attributions tended to be *lower* than moral responsibility attributions, even when the very same participants

answered both questions. I still collected participants' free will judgments for those interested, but I won't comment on them.

2) The Deep Self measures I will be using are slightly different from those used in the literature. They have been rephrased to avoid all references to *concordance* between action and deep self and to rather insist on the *provenance* of one's action from one's deep self. Indeed, my own spin on the Deep Self hypothesis (i.e. the Deep Self Provenance hypothesis) is that one is responsible for actions that stem from one's Deep Self, even if there is no concordance between the final outcome and one's Deep Self (see Cova, 2011).
3) I am not committed to the claim that the effect of manipulation on attributions of moral responsibility will be *fully* mediated by participants' scores on Deep Self measures. Because Sripada's Deep Self account (Sripada, 2010) has been tested using mediation analysis, some have argued that the Deep Self explanation of intuitions about manipulation argument fails because Deep Self scores only *partially* mediate the effect of Deep Self on moral responsibility (Björnsson, 2016). However, we should expect Deep Self scores to *fully* mediate the effect of manipulation only in case our measures of Deep Self beliefs and attributions of moral responsibility are accurate enough. Using simulations, I have observed that introducing some measurement error prevents Deep Self scores from fully mediating the effect of manipulation, even when attributions of responsibility are fully driven by considerations about the Deep Self. Given that Deep Self measures are phrased in metaphoric terms and that we should expect participants not to have full access to their internal representation of such cases, expecting full mediation seems an unreasonable demand.
4) Rather than focusing only on mediation analysis, I prefer to put the Deep Self Concordance model to the test by showing that it makes very specific and novel predictions and testing these predictions. In this paper, this prediction is the following: *the effect of manipulation on attributions of moral responsibility should be lower for good actions than for bad actions*. Indeed, let's suppose that the reason why some people consider agents in manipulation cases not to be morally responsible is because they perceive them as not acting from their own true, deeply-held values. The effect of manipulation should be lower when it is *easier* to think that the action performed comes from one's true values. However, a vast literature has documented the following fact: people find it easier to think that people's deep selves are *morally good* than to consider that people can be 'rotten to the core' (Newman et al., 2014; Strohminger et al., 2017). Thus, we should expect people to find it easier to believe that a manipulated agent acts upon one's true values when these values are morally good.

In the next section, I present results suggesting that this bold hypothesis actually works in the case of classical manipulation cases, thus suggesting that the Deep Self Provenance model is a good explanation of participants' intuitions about manipulation cases. Then, in Studies 1 and 2, I extend this hypothesis to Zygote cases.

5. Materials and data

Materials and data for all studies are available at osf.io/qnp86/

6. Pilot study - The effect of manipulation on moral responsibility depends on the action's moral valence

In this pilot study, the goal was simply to test one key prediction of my Deep Self Provenance model: that the effect of manipulation on judgments of moral responsibility (when comparing a manipulation case to a control case) should be greater for bad actions, compared to good actions.

6.1. Materials and Methods

The study took the form of an online survey. Participants were presented with one of four vignettes. Two vignettes were Sripada's *Modified* and *Control* vignettes. The two others were modified versions of the *Modified* and *Control* vignette, in which the agents' action was morally good (saving Mrs. White) rather than morally bad (killing Mrs. White).[2] Thus, two dimensions were independently manipulated: whether the agent was manipulated, and his action's moral valence.

After reading the vignette, participants were presented with a series of statements. For each of them, participants were asked to indicate to which extent they agreed (-3: Strongly disagree, 3 = Strongly agree):

> *(Resp.)* Bill is morally responsible for killing [saving] Mrs. White. (Participants were asked to justify their answer.)
> *(Blame/Praise)* Bill deserves blame for killing [praise for saving] Mrs. White.
> *(Desert)* Bill deserves to be punished for killing [rewarded for saving] Mrs. White.
> *(Free Will)* Bill killed [saved] Mrs. White of his own free will.
> *(Deep Self 1)* It is Bill's very own desires and values that led him to kill [save] Mrs. White.
> *(Deep Self 2)* It's what Bill really wanted deep down that caused him to kill [save] Mrs. White.
> *(Deep Self 3)* It is because of the kind of person he's truly deep down inside that Bill killed [saved] Mrs. White.
> *(Deep Self 4)* Bill's killing [saving] of Mrs. White had nothing to do with what he really wanted to do.
> *(Bypass 1)* Bill's own decisions played no role in his killing [saving] Mrs. White.
> *(Bypass 2)* What Bill wanted had no effect on his killing [saving] Mrs. White.
> *(Bypass 3)* What Bill believed had no effect on his ending up to kill [save] Mrs. White.
> *(Control)* Bill had no control over his killing [saving] Mrs. White.

[2] Originally, I also planned to add two versions of Sripada's *Manipulation* case. However, due to human error, the Good version of the *Control* case was inserted in place of the Good version of the *Manipulation* case. I thus excluded the *Manipulation* cases from analysis.

> *(Throughpass)* When earlier events caused Bill's actions, they did so by affecting what he believed and wanted, which in turn caused him to act in a certain way.

After that, participants answered four comprehension checks and completed a scale (the Geneva Sentimentality Scale; Cova & Boudesseul, 2021) in which two attention checks were hidden.

6.2. Results

491 participants recruited on Prolific Academic and paid £0.70 for their participation completed our survey. After exclusion based on 4 comprehension checks and 2 attention checks, we were left with 391 participants (235 men, 152 women and 4 others; $M_{\text{age}} = 25.73$, $SD_{\text{age}} = 7.82$).

Responsibility ratings. Responsibility ratings, blame/praise ratings and desert ratings were aggregated to form a single aggregated responsibility score (ARS; α = .72). A 2-way ANOVA with Outcome Valence (Bad/Good) and Case (Modified vs. Control) as factors and ARS as dependent variable found a significant interaction effect: $F(1,387) = 33.86$, $p < .001$, $\eta_\rho^2 = 0.08$, a significant main effect of Outcome Valence: $F(1,387) = 45.14$, $p < .001$, $\eta_\rho^2 = 0.09$, and a significant main effect of Case: $F(1,387) = 46.99$, $p < .001$, $\eta_\rho^2 = 0.11$. Results are presented in Table 2.

	Bad		Good	
	Modified	Control	Modified	Control
ARS	1.29*** (1.33) 82%	2.61 (0.66) 98%	1.31 (1.04) 86%	1.44 (0.99) 89%
Resp.	1.03*** (1.62) 75%	2.68 (0.73) 98%	1.12* (1.71) 72%	1.61 (1.52) 84%
Blame/Praise	1.19*** (1.61) 76%	2.53 (0.92) 95%	1.60 (1.32) 81%	1.89 (1.13) 89%
Desert	1.65*** (1.31) 87%	2.62 (0.76) 96%	1.21 (1.32) 69%	0.82 (1.53) 61%
Free Will	0.65*** (1.77) 65%	2.52 (1.14) 95%	0.88*** (1.58) 65%	2.38 (0.91) 96%
Deep Self	-0.09*** (0.98) 47%	0.64 (0.93) 74%	0.42*** (0.82) 65%	0.90 (0.59) 91%
Bypass	-1.05*** (1.09) 11%	-1.78 (1.02) 4%	-0.97* (1.14) 17%	-1.30 (1.16) 10%

Control	-0.96*** (1.69) 24%	-2.29 (0.96) 1%	-0.94*** (1.62) 22%	-1.94 (1.17) 5%
Throughpass	1.39*** (1.17) 82%	-0.20 (1.77) 35%	0.89*** (1.59) 63%	0.21 (1.20) 46%
N	79	112	98	102

Table 2. Mean, Standard Deviation and % of answers above the midpoint (= 0) for each condition in Pilot Study. Stars(*) present the results of Welch t-tests comparing the two cases (Modified, and Control) for each variable and each outcome valence. Interaction between Outcome Valence and Case was significant for: ARS, Resp, Blame/Praise, Desert, and Throughpass; but not significant for: Free Will, Deep Self, Bypass, and Control.

6.3. Discussion

As predicted by the Deep Self Provenance hypothesis, the effect of manipulation was stronger for bad actions than for good actions. This is explained naturally by the fact that people are less likely to think that someone committing a bad action acted in accordance with their Deep Self.

Thus, if people's intuitions about Zygote cases are explained by people's intuitions about the Deep Self, we should expect to observe the same pattern. This is what I set out to test in Study 1.

7. Study 1 - Investigating intuitions about the Zygote Argument

7.1. Materials and Methods

The study took the form of an online survey. At the beginning of the experiment, participants were randomly assigned to one of three cases (Manipulation, Indeterministic or Control) and to one of two outcome valences (Bad or Good). Then, depending on these factors, each participant was presented with one of six cases.

Manipulation	*Indeterministic*	*Control*
Imagine a world just like ours, except that there exist beings similar to the Greek and Roman gods. These beings are not all-powerful but they have strange powers: they can see into the future to determine with certainty what consequences their actions will have. They can also manipulate matter and even create life. These beings' existence is completely unknown to men.	Imagine a world just like ours, except that there exist beings similar to the Greek and Roman gods. These beings are not all-powerful but they have strange powers: they can see into the future to determine with great accuracy what consequences their actions will have. They can also manipulate matter and even create life. These beings' existence is completely unknown to men.	Imagine a world just like ours, except that there exist beings similar to the Greek and Roman gods. These beings are not all-powerful but they have strange powers: they can see into the future to determine with great accuracy what consequences their actions will have. They can also manipulate matter and even create life. These beings' existence is completely unknown to men.

Imagine that Diane is such a goddess. She knows with full certainty that, if she creates a certain zygote (a human embryo at a very early stage) and implants it in the uterus of Mary, a common human, on March 3rd, 2021 there is a 100% chance that this zygote will grow up to become a young man named Bill. Diane also sees that there is a 100% chance that, on his 30th birthday, Bill will eventually **kill his aunt to inherit from her early**. Knowing all that Diane creates said zygote and implants it in Mary's uterus. After that, she never interferes again in the zygote's development and Bill's life. Nine months later, Mary gives birth to Bill. Bill grows up to be an ordinary human being. He's as rational as other human beings and exerts as much self-control upon his actions. On his 30th birthday, after a long deliberation, and on the basis of his deeply-held values, **he kills his aunt to inherit from her early**.	Imagine that Diane is such a goddess. She knows with full certainty that, if she creates a certain zygote (a human embryo at a very early stage) and implants it in the uterus of Mary, a common human, on March 3rd, 2021 there is a 98% chance that this zygote will grow up to become a young man named Bill. Diane also sees that there is a 95% chance that, on his 30th birthday, Bill will eventually **kill his aunt to inherit from her early**. Knowing all that Diane creates said zygote and implants it in Mary's uterus. After that, she never interferes again in the zygote's development and Bill's life. Nine months later, Mary gives birth to Bill. Bill grows up to be an ordinary human being. He's as rational as other human beings and exerts as much self-control upon his actions. On his 30th birthday, after a long deliberation, and on the basis of his deeply-held values, **he kills his aunt to inherit from her early**.	Imagine that Diane is such a goddess. She knows with full certainty that, if she creates a certain zygote (a human embryo at a very early stage) and implants it in the uterus of Mary, a common human, on March 3rd, 2021 there is a 98% chance that this zygote will grow up to become a young man named Ted. Diane also sees that there is a 95% chance that, on his 30th birthday, **Ted will rob a bank**. Knowing all that, Diane decides not to create said zygote and never interferes in Mary's life. A few days later, Mary has sex with a man named John and gets pregnant. Nine months later, Mary gives birth to Bill. Bill grows up to be an ordinary human being. He's as rational as other human beings and exerts as much self-control upon his actions. On his 30th birthday, after a long deliberation, and on the basis of his deeply-held values, **he kills his aunt to inherit from her early**.

Table 3. Three different cases used in Study 1 (Bad Outcome version).

The three different cases in their Bad Outcome version are presented in Table 3. The Manipulation case is supposed to be a typical Zygote case: the agent is created by a goddess who predicts his action with a 100% chance. The Indeterministic case is similar, except that the goddess cannot perfectly predict the future. Finally, the Control case is a case in which the goddess still predicts the agent's future with near certainty, but in which she does not play a role in the production of this agent (rather, he is the normal product of human sexual intercourse).

I introduced the Indeterministic case because the Zygote argument assumes that determinism plays a key role in the intuition that the agent is not morally responsible in Zygote cases. However, if it turned out that people also had the intuition that agents still lack moral responsibility in modified Zygote cases in which the goddess' actions only have indeterministic effects, this would pose a problem to both versions of the Zygote argument. On the *no difference* version of the Zygote argument, this would lead us to conclude that moral responsibility is not only impossible in a deterministic universe, but also in indeterministic ones. Thus, we would not conclude to incompatibilism, but to some form of skepticism about moral responsibility. On the *explanation-based* version of the Zygote argument, this would lead us to conclude that determinism is not the reason why agents are not morally responsible in Zygote cases, and this would prevent us from concluding that determinism is incompatible with responsibility.[3]

In the Good outcome version of these vignettes, Bill's bad action (indicated in bold in Table X) was replaced by a morally good one ('giving the money he just inherited from his aunt to create a homeless shelter'). In the Control case, 'robbing a bank' was also replaced by 'saving someone from drowning').

After reading the vignette, participants were presented with roughly the same statements as in the Pilot Study, except that the action description was replaced by "killing his aunt" (in the Bad Outcome version) or "creating a homeless shelter" (in the Good Outcome version). The only statements to be modified more extensively were the following:

> *(Resp)* It would be fair to hold Bill morally responsible for having killed his aunt [created a homeless shelter].[4]
> *(Blame/Praise)* Bill deserves blame [praise] for killing his aunt [creating a homeless shelter].
> *(Desert)* Bill deserves to be punished [deserves credit] for killing his aunt [creating a homeless shelter].

7.2. Research questions and hypotheses

In this study and the next one, the questions I wanted to address were the following:

1. Do people really have the intuition that agents in Zygote cases are not morally responsible for their actions?
2. Does manipulation affect participants' attributions of moral responsibility?
3. Does the effect of manipulation depend on determinism? Or is manipulation without determinism enough to trigger the same intuitions?
4. Are agents in manipulation cases identical to agents in control cases? Or do people perceive differences between the two that could explain their different assessment of their moral responsibility?
5. Can the results be explained by the Deep Self Provenance hypothesis? Which can be broken down in three sub-questions:
    a. are attributions of moral responsibility correlated with Deep Self scores?
    b. do Deep Self scores mediate the effect of manipulation on attribution of moral responsibility?
    c. does the effect of manipulation depend on the action's moral valence?

[3] For the idea that the Zygote argument fails if determinism plays no role in the intuition that agents lack moral responsibility, see Kearns (2012).

[4] This cumbersome formulation was chosen because analysis of participants' open-ended justifications in Pilot Study revealed that, in the Good Outcome cases, a lot of participants interpreted the question about whether Bill was responsible for saving Mrs. White as being about whether it was Bill's duty and obligation to save Mrs. White. This ambiguity in the expression "being morally responsible for" is a serious problem and has already plagued other studies (such as Turri, 2017a, 2017b).

7.3. Results

705 participants recruited on Prolific Academic and paid £0.75 for their participation completed our survey. After exclusion based on 4 comprehension checks and 2 attention checks, I was left with 532 participants (272 women, 246 men and 14 others; $M_{\text{age}} = 31.38$, $SD_{\text{age}} = 11.74$).

Q1. Do people have the intuition that agents in Zygote cases are not morally responsible for their actions? Responsibility ratings, blame/praise ratings and desert ratings were aggregated to form a single aggregated responsibility score (ARS; $\alpha = .83$). Responsibility scores in the Manipulation condition were high ($M = 1.79$, $SD = 1.35$) and significantly above the midpoint: $t(147) = 16.16$, $p < .001$. Overall, 88% of participants obtained ARS superior to 0. Thus, it seems fair to conclude that most participants did not share the intuition that agents in Zygote cases were not morally responsible for their actions.

Q2. Does manipulation affect participants' attributions of moral responsibility? I conducted a 2-way ANOVA with Outcome Valence (Bad/Good) and Case (Control/Manipulation/Indeterministic) as factors and ARS as dependent variable. We found a main effect of Outcome Valence: $F(1,526) = 37.01$, $p < .001$, $\eta_\rho^2 = 0.06$, a main effect of Case: $F(2,526) = 11.48$, $p < .001$, $\eta_\rho^2 = 0.04$, and a significant interaction effect: $F(2,526) = 5.29$, $p = .005$, $\eta_\rho^2 = 0.02$. Thus, manipulation had an effect on aggregated responsibility ratings, but this effect depended on outcome valence (i.e. whether the action was good or bad). Thus, I analyzed bad outcomes cases and good outcomes separately, by performing two separate Tukey tests. As summarized in Table 4, I found no effect of manipulation for good outcome cases. However, for bad outcome cases, I found that scores in the Control condition were significantly higher compared to the Manipulation and Indeterministic condition. Thus, for bad outcomes only, manipulation had an effect on participants' ARS.

	Bad outcome			*Good outcome*		
	Manip.	*Indet.*	*Control*	*Manip.*	*Indet.*	*Control.*
ARS	1.90^{a} (1.51) 86%	2.23^{a} (1.03) 96%	2.73^{b} (0.63) 98%	1.66 (1.14) 91%	1.87 (0.86) 97%	1.82 (1.01) 90%
Resp.	1.76^{a} (1.79) 84%	2.20^{a} (1.11) 94%	2.75^{b} (0.88) 97%	0.91 (1.83) 63%	1.34 (1.66) 77%	1.47 (1.54) 73%
Blame/Praise	1.74^{a} (1.83) 81%	2.28^{b} (1.03) 94%	2.76^{c} (0.71) 96%	2.01 (1.13) 93%	2.10 (0.82) 95%	1.83 (1.05) 85%
Desert	2.21^{a} (1.34) 88%	2.21^{a} (1.07) 93%	2.67^{b} (0.73) 97%	2.06 (1.21) 91%	2.19 (0.77) 97%	2.14 (1.10) 92%

Free will	1.60^{a} (1.89) 79%	2.00^{a} (1.26) 91%	2.72^{b} (0.67) 97%	1.34^{a} (1.79) 75%	1.79^{a} (1.42) 82%	2.27^{b} (1.09) 92%
Deep Self	1.29^{a} (1.30) 83%	1.53^{a} (1.13) 90%	1.94^{b} (0.94) 94%	1.45^{a} (1.38) 85%	1.72^{ab} (1.10) 92%	1.90^{b} (0.89) 94%
Bypass	-1.25^{a} (1.38) 15%	-1.84^{b} (1.13) 06%	-2.06^{b} (1.03) 05%	-1.30 (1.39) 13%	-1.56 (1.16) 08%	-1.75 (1.08) 04%
Control	-1.21^{a} (1.89) 18%	-2.10^{b} (1.30) 06%	-2.66^{c} (0.70) 01%	-1.25^{a} (1.87) 19%	-1.63^{ab} (1.47) 10%	-2.17^{b} (1.21) 05%
Throughpass	0.44 (1.72) 50%	0.46 (1.65) 54%	0.20 (2.02) 50%	0.44 (1.97) 53%	0.48 (1.79) 49%	0.28 (1.67) 49%

Table 4. Mean, Standard Deviation and % of answers above the midpoint (= 0) for each condition in Study 1. Superscripts present the results of Tukey tests comparing the three cases (Manipulation, Indeterministic, and Control) for each variable and each outcome valence. When superscripts are present, two conditions that do not share a common letter in superscript significantly differ from each other. When no superscript is present, this means that there was no difference between conditions. Interaction between Outcome Valence and Condition was significant for: ARS, Blame/Praise, and Desert ; but not significant for: Resp., Free Will, Deep Self, Bypass, Control, and Throughpass.

Q3. Does the effect of manipulation depend on determinism? Focusing on bad outcome cases, I found a significant difference in ARS between the Indeterministic and the Control conditions, but no significant difference between the Manipulation and Indeterministic conditions. Thus, it seems that manipulation can have an effect on participants' attributions of responsibility even in absence of determinism.

Q4. Are agents in manipulation cases identical to agents in control cases? Focusing on bad outcome cases, I used two Tukey tests to compare Deep Self scores ($\alpha = .82$) and Bypassing scores ($\alpha = .73$) across conditions. Both tests found significant differences between the Manipulation and Control cases (see Table 4): Deep Self scores were lower in the Manipulation condition while Bypassing scores were higher in the Manipulation condition. This suggests that participants did not perceive agents in the Manipulation condition as identical to those in the Control condition.

Q5. Can the results be explained by the Deep Self Provenance hypothesis?

(a) Interaction between condition and valence: As predicted by the Deep Self Provenance hypothesis, the effect of manipulation was stronger for bad outcomes than for good outcomes.

(b) Correlations: There was a significant correlation ($r = .60$) between ARS and Deep Self scores (see Table 5).

	Free Will	Deep Self	Bypass	Control
ARS	.74***	.60***	-.53***	-.63***
Free Will	-	.67***	-.55***	-.68***
Deep Self	-	-	-.65***	.64***
Bypass	-	-	-	.68***

Table 5. Inter-correlations between variables in Study 1.

(c) Mediation analysis: I used structural equation modelling (in R's {lavaan} package) to investigate whether Deep Self scores mediated the effect of manipulation on ARS. I focused on bad outcome cases, and on the two following comparisons: Manipulation vs. Control and Indeterministic vs. Control. Results are presented in Figure 1. In both cases, the effect of manipulation on ARS was significantly mediated by Deep Self ratings. However, in both cases, a direct effect remained.

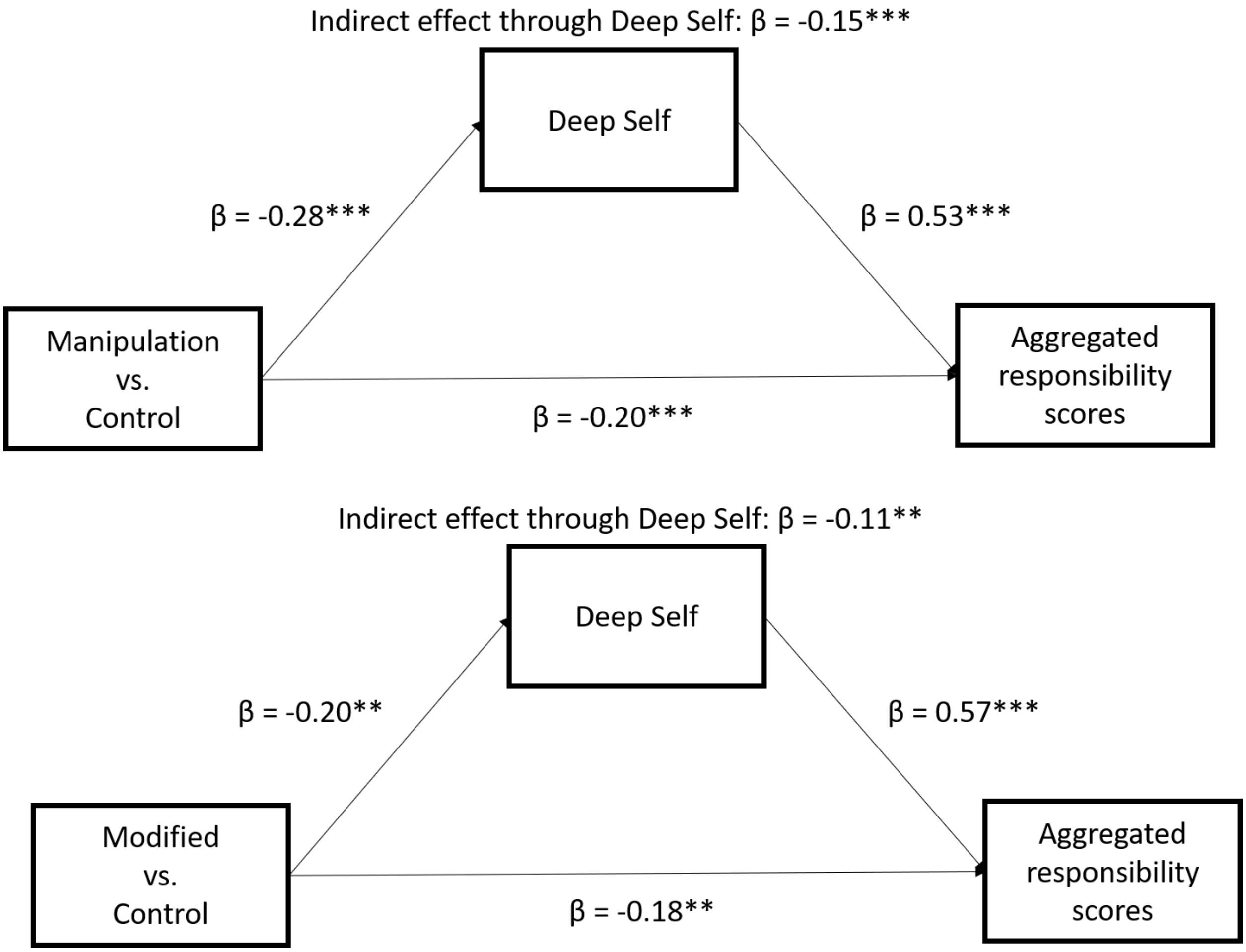
Indirect effect through Deep Self: β = -0.15***
Deep Self
β = -0.28***
β = 0.53***
Manipulation vs. Control
Aggregated responsibility scores
β = -0.20***
Indirect effect through Deep Self: β = -0.11**
Deep Self
β = -0.20**
β = 0.57***
Modified vs. Control
Aggregated responsibility scores
β = -0.18**

Figure 1. Deep Self scores as mediators of the effect of manipulation (Manipulation vs. Control and Modified vs. Control) in Study 1.

Thus, my three sub-hypothesis were corroborated by my data. Does this mean that participants' judgments of responsibility are best explained by participants' perception of agents' deep self? Not necessarily, as there is also some contradictory evidence. For example, in positive outcome cases, there was no difference in responsibility judgments between the Manipulation and Control cases, but there was a significant difference in Deep Self scores.

7.4. Discussion

In this study, I investigated participants' intuitions about Zygote cases. The results of this study suggest the following conclusions:

1. The results suggest both a soft- and a hard-line answer to the Zygote argument. On the hard-line side, most participants actually had the intuition that agents were morally responsible for their actions in the Zygote cases. On the soft-line side, even the answers of the minority who considered that agents in Zygote cases were not morally responsible can

be explained away by the fact that agents in the Zygote cases are not perceived as identical to participants in the Control cases.

2. Our results suggest that at least part of the effect of manipulation in Zygote cases has nothing to do with determinism, as there was also a difference between the Manipulation and the Control case.

Overall, this suggests that laypeople's intuitions about Zygote cases do not support Mele's argument. Nevertheless, one might object that my formulation of the Manipulation case failed to elicit the appropriate intuitions because it failed to highlight the relevant considerations. For example, maybe it did not make clear the fact that Diane's ability to predict the future was premised on some form of determinism.

Another objection might be that, because I used particularly bad and good deeds (killing a person vs. giving all one's money to charity), their desire to blame (or praise) agents might have led participants to ignore other relevant features of these scenarios.

In the next study, I modified my presentation of the Zygote argument to address these issues.

8. Study 2 - Replication and extension of previous findings

8.1. Materials and Methods

In this study, the three types of vignettes used in Study 1 were modified to make it more salient that the goddess was able to predict the future because she has perfect knowledge of the current state of the world and of the laws of nature. The probabilities in the Indeterministic case were modified: the goddess now saw that the zygote had a 99% chance to grow up and become Bill and that Bill had a 67% chance of performing the target action. The Control case was also rewritten to exclude all reference to Ted and to simply be a case in which the goddess predicts (with a 67% chance) that Bill would perform the target action, without her having anything to do with Bill's birth. A fourth type of vignette was included, in which the goddess did not create Bill but was able to predict his future behavior with perfect certainty. This way, I was able to manipulate Manipulation (M) and Determinism (D) orthogonally by having four cases: Manipulation (M+/D+), Indeterministic (M+/D-), Deterministic (M-/D+) and Control (M-/D-). This would allow me to determine the respective weight of Manipulation and Determinism in people's intuitions about Zygote cases.

Compared to Study 1, I also modified the type of actions performed by Bill, so that they would be less upsetting and more closely matched across conditions. In the Bad Outcome version, Bill finds a wallet full of money and keeps it for himself. In the Good Outcome version, Bill returns the wallet to its rightful owner.

As an example, here is the Good version of the Deterministic case:

> *Deterministic (M-/D+), Good:* Imagine a world just like ours, except that there exist beings similar to the Greek and Roman gods. These beings are not all-powerful but they have strange powers: they can use their perfect knowledge of what is currently happening in the world and combine it with their perfect knowledge of the laws of nature to deduce with absolute certainty what will happen in the future and what consequences their actions will have. They can also manipulate matter and even create and destroy life. These beings' existence is completely unknown to men.
>
> Imagine that Diane is such a goddess. One day, March 3rd, 2021, she notices that Mary, a common human, has just got pregnant. Based on her knowledge of the current state of the world and the laws of nature, she knows with full certainty that, if she leaves the zygote (a human embryo at a very early stage) in Mary's uterus grow without interfering, there is a 100% chance that this zygote will grow up to become a young man named Bill. Diane also sees every action that Bill will perform in his life. For example, she sees that there is a 100% chance that, on his 30th birthday, Bill will find a wallet containing $500 in the street and that he will decide not to keep the money for himself but to return it to its rightful owner.
>
> Knowing all that Diane decides not to interfere and to leave the zygote grow in Mary's uterus. Moreover, she never interferes again in the zygote's development and Bill's life.
>
> Nine months later, Mary gives birth to Bill. Bill grows up to be an ordinary human being. He's as rational as other human beings and exerts as much self-control upon his actions. On his 30th birthday, he finds a wallet containing $500 in the street. After a long deliberation, and on the basis of his deeply-held values, he decides not to keep the money for himself but to return it to its rightful owner.

After reading the vignette, participants were asked to rate the same statements as in Study 1. The *Throughpass* statement was removed as it did not seem to track exactly what it was supposed to be tracking.[5] The only statement to be modified beyond action description was:

> (*Resp*) Bill is the one who is morally responsible for having kept [returned] the money.

7.2. Results

1350 participants recruited on Prolific Academic and paid £0.63 for their participation completed our survey. After exclusion based on 3 comprehension checks[6], I was left with 1111 participants (611 women, 483 men and 17 others; $M_{\text{age}} = 32.93$, $SD_{\text{age}} = 11.70$).

[5] Introduced by Björnsson and Pereboom, D. (2014), this item was supposed to measure the belief that "the agent's deliberation is not bypassed". However, the results of the Pilot Study show that Throughpass scores are way higher in the manipulation than in the Control case, suggesting that the statement rather measures to which extent the agent is manipulated by external forces. For further discussion, see Cova (forthcoming).

[6] There were actually 4 comprehension checks, but the last one was failed by a lot of participants. Comments sent spontaneously by participants on Prolific Academic indicated that they found the sentence ill-phrased and that they were unsure what to answer. The sentence (that they had to rate as TRUE or FALSE) was "Diane predicted that there would be a 100% chance that the zygote would grow up to give birth to Bill". According to a lot of participants, it was not correct to say that the zygote "gave birth" to Bill.

Q1. Do people have the intuition that agents in Zygote cases are not morally responsible for their actions? Responsibility ratings, blame/praise ratings and desert ratings were aggregated to form a single aggregated responsibility score (ARS; α = .71). Responsibility scores in the M+/D+ condition were high (M = 1.25, SD = 1.36) and significantly above the midpoint: t(301) = 15.95, p < .001. Overall, 78% of participants in the M+/D+ condition obtained an ARS superior to 0. Thus, it seems fair to conclude that most participants did not share the intuition that agents in Zygote cases were not morally responsible for their actions.

Q2. Does manipulation affect participants' attributions of moral responsibility? I conducted a 2-way ANOVA with Outcome Valence (Bad/Good) and Case (M+/D+, M+/D-, M-/D+ and M-/D-) as factors and ARS as dependent variable. There was no significant interaction effect: F(3,1103) = 1.24, p = .29, η_{ρ}^{2} = 0.003. After dropping the interaction term, we found a significant effect of Valence: F(1,1106) = 219.71, p < .001, η_{ρ}^{2} = 0.15, but no significant effect of Case: F(3,1106) = 1.98, p = .12, η_{ρ}^{2} = 0.005. Thus, there was no difference between the Manipulation (M+/D+) and Control (M-/D-) cases, even when analyzing both outcome valences separately.

	Valence	Case			
		Manipulation (M+/D+)	Indet. (M+/D-)	Deterministic (M-/D+)	Control (M-/D-)
ARS	Bad	0.85 (1.43)	0.99 (1.18)	0.92 (1.29)	1.11 (1.20)
	Good	1.82 (1.06)	1.80 (0.88)	2.09 (0.87)	1.99 (0.87)
Resp.	Bad	1.48 (1.92)	1.87 (1.28)	1.83 (1.45)	1.93 (1.46)
	Good	2.08^{a} (1.14)	2.08^{a} (1.04)	2.49^{b} (0.77)	2.46^{b} (0.74)
Blame/Praise	Bad	1.04 (1.80)	1.02 (1.69)	1.05 (1.76)	1.32 (1.59)
	Good	1.55 (1.39)	1.50 (1.25)	1.84 (1.24)	1.65 (1.27)
Desert	Bad	0.02 (1.77)	0.09 (1.61)	-0.13 (1.58)	0.07 (1.62)
	Good	1.83 (1.32)	1.83 (1.15)	1.93 (1.05)	1.86 (1.27)
Free Will	Bad	1.98 (1.44)	2.10 (1.14)	2.12 (1.24)	2.24 (1.11)

	Good	1.74^{a} (1.43)	2.03^{ab} (1.10)	2.29^{bc} (1.06)	2.41^{c} (0.89)
Deep Self	Bad	1.46 (1.17)	1.56 (1.02)	1.42 (1.09)	1.59 (1.04)
	Good	1.51^{a} (1.15)	1.60^{ab} (0.96)	1.86^{b} (0.87)	1.87^{b} (0.82)
Bypass	Bad	-1.51 (1.26)	-1.56 (1.13)	-1.49 (1.20)	-1.60 (1.18)
	Good	-1.41^{a} (1.18)	-1.14^{b} (1.33)	-1.46^{ab} (1.24)	-1.53^{ab} (1.20)
Control	Bad	-1.86 (1.46)	-2.05 (1.08)	-1.95 (1.21)	-2.17 (1.21)
	Good	-1.75 (1.47)	-1.91 (1.11)	-2.22 (0.97)	-2.29 (0.99)
N	Bad	139	129	119	115
	Good	118	115	116	260

Table 6. Mean and Standard Deviation for each condition in Study 2. Superscripts present the results of Tukey tests comparing the four cases for each variable and each outcome valence. When superscripts are present, two conditions that do not share a common letter in superscript significantly differ from each other. When no superscript is present, this means that there was no difference between conditions. There was no significant interaction effect.

	Free Will	Deep Self	Bypass	Control
ARS	.44***	.46***	-0.28***	-0.36***
Free Will	-	.55***	-.41***	-.64***
Deep Self	-	-	-.57***	-.60***
Bypass	-	-	-	.57***

Table 7. Inter-correlations between variables in Study 2.

In absence of a significant main effect of Case and/or of a significant interaction effect between Cases and Outcome Valence, it was impossible to investigate my other research questions. This is why I decided to run additional data.

7.4. Additional data

Aggregate responsibility scores were abnormally low in the Control (M-/D-) condition, particularly when the outcome was bad (M = 1.11). This might simply have been because participants did not consider keeping the wallet as a serious offense worth blaming and punishing. But it might also have been because this condition was not an adequate control condition - maybe the mere presence of superior entities able to (roughly) predict and prevent human action was already perceived as a threat to humans' moral responsibility. After all, given that our goddess could have prevented Bill from keeping the wallet, why not think *she* is partly responsible for it?

To adjudicate between these two possibilities, I decided to collect more data by introducing two *pure control* cases (that I called *normal* cases), in which there is absolutely no reference to any god or goddess (see Table 8).

Normal (Bad)	*Normal (Good)*
Imagine a world just like ours. One day, March 3rd, 2021, Mary, a common human, gets pregnant. The zygote (a human embryo at a very early stage) in Mary's uterus grows up to become a young man named Bill. On his 30th birthday, Bill finds a wallet containing $500 in the street. After a long deliberation, and on the basis of his deeply-held values, he decides to keep the money for himself rather than return it to its rightful owner.	Imagine a world just like ours. One day, March 3rd, 2021, Mary, a common human, gets pregnant. The zygote (a human embryo at a very early stage) in Mary's uterus grows up to become a young man named Bill. On his 30th birthday, Bill finds a wallet containing $500 in the street. After a long deliberation, and on the basis of his deeply-held values, he decides not to keep the money for himself but to return it to its rightful owner.

Table 8. The two *Normal* cases used in Study 2 to collect additional data.

I recruited 300 additional participants on the same online platform, using the same demographic constraint, and launching the additional study at the same hour I launched the original study. Based on three comprehension checks, 45 participants were excluded, leaving us with 255 participants (133 women, 116 men and 6 others; M_{age} = 33.00, SD_{age} = 12.26).

Participants received one of our two normal cases at random, then answered the same questions as in Study 2 (only comprehension checks were changed). Then, I compared participants' answer to these normal cases with participants' answer to the Manipulation (M+/D+) cases in Study 2.

For ARS, an ANOVA with Outcome Valence (Bad/Good) and Case (Manipulation vs. Normal) as factors revealed a marginally significant interaction effect: $F(1,508) = 3.13$, $p = .078$, $\eta_{\rho}^2 = 0.006$, a significant main effect of Outcome Valence: $F(1,508) = 64.84$, $p < .001$, $\eta_{\rho}^2 = 0.109$, and a significant main effect of Case: $F(1,508) = 5.06$, $p = .025$, $\eta_{\rho}^2 = 0.010$.

	Bad		Good	
	Manipulation	Normal	Manipulation	Normal

ARS	0.85* (1.43)	1.25 (1.09)	1.82 (1.06)	1.87 (0.89)
Resp.	1.48* (1.92)	1.98 (1.59)	2.08 (1.14)	2.32 (0.90)
Blame/Praise	1.04** (1.79)	1.55 (1.36)	1.55 (1.39)	1.60 (1.25)
Desert	0.02 (1.77)	0.22 (1.56)	1.83 (1.32)	1.68 (1.25)
Free Will	1.98*** (1.44)	2.51 (0.95)	1.74*** (1.43)	2.42 (0.71)
Deep Self	1.46* (1.17)	1.75 (0.93)	1.51 (1.15)	1.58 (0.85)
Bypass	-1.51*** (1.26)	-2.00 (0.93)	-1.41 (1.18)	-1.57 (1.17)
Control	-1.86*** (1.46)	-2.51 (0.73)	-1.75** (1.47)	-2.28 (1.08)
N	139	122	118	133

Table 9. Mean and Standard Deviation for each condition in follow-up to Study 2. Stars(*) present the results of Welch t-tests comparing the two cases (Manipulation and Normal) for each variable and each outcome valence.

So, using these *Normal* cases as comparison cases for our Zygote cases, I was able to find some effect of manipulation on participant's judgments about moral responsibility. This allowed me to investigate my research questions further, by focusing on Bad Outcome cases, as there was no significant difference in ARS between the two cases for Good Outcome cases.

Q3. Does the effect of manipulation depend on determinism? As we saw earlier, for Bad Outcome cases, there was no significant difference in ARS between the M+/D+ and M+/D- cases. Thus, it is not clear that Determinism plays a role in undermining agents' moral responsibility in Zygote cases.

Q4. Are agents in manipulation cases identical to agents in control cases? Focusing on bad outcome cases, I used two Welch t-tests to compare Deep Self scores (α = .82) and Bypassing scores (α = .73) across conditions. Both tests found significant differences between the Manipulation and Normal cases (see Table 9): Deep Self scores were lower in the Manipulation condition while Bypassing scores were higher in the Manipulation condition. This suggests that participants did not perceive agents in the Manipulation condition as identical to those in the Normal condition.

Q5. Can the results be explained by the Deep Self Provenance hypothesis?

(a) Interaction between condition and valence: There was a marginally significant interaction effect. Moreover, the effect of manipulation was significant for bad outcomes but not for good outcomes. This in line with my predictions.

(b) Correlations: For bad outcomes, there was a significant correlation (*r* = .43) between ARS and Deep Self scores.

c) Mediation analysis: I used structural equation modelling (in R's {lavaan} package) to investigate whether Deep Self scores mediated the effect of manipulation on ARS. I focused on bad outcome cases, and on the comparison between the Manipulation and Normal cases. Results are presented in Figure 2. The effect of manipulation on ARS was significantly mediated by Deep Self ratings. No significant direct effect remained.

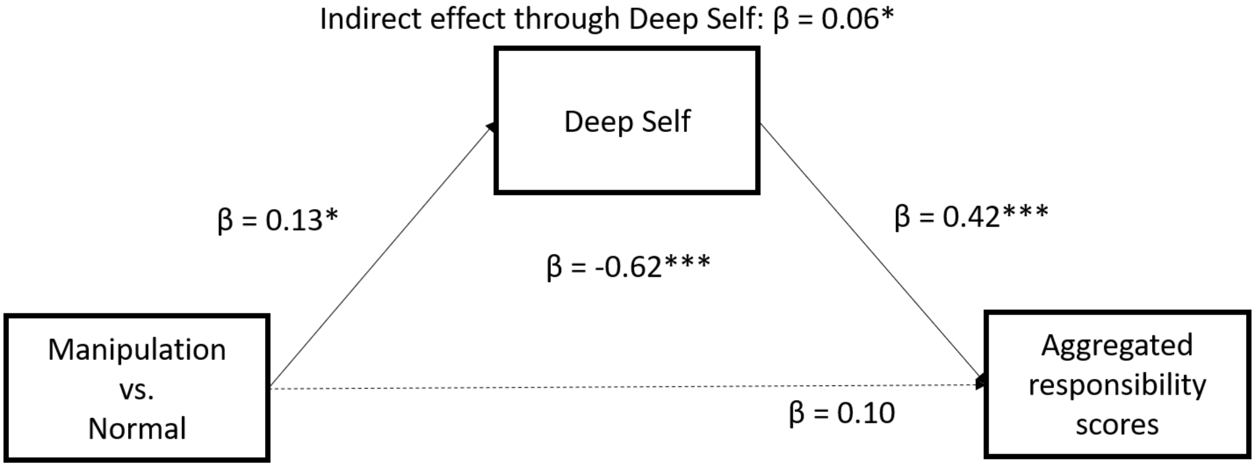
Indirect effect through Deep Self: β = 0.06*
Deep Self
β = 0.13*
β = 0.42***
β = -0.62***
Manipulation vs. Normal
Aggregated responsibility scores
β = 0.10

Figure 2 Deep Self scores as mediators of the effect of manipulation (Manipulation vs. Normal) in follow-up to Study 2.

8. Conclusion: "the Goddess you need can't be me"

In this paper, my goal was to investigate people's intuitions about the Zygote argument, and to see whether they supported the argument. Overall, my data did not really support the Zygote argument, as it allowed for both a *hard-line* and *soft-line* response to the Zygote argument.

Regarding the *hard-line* response, it is striking to observe that most participants did attribute moral responsibility to agents in Zygote cases. When the outcome was good, there was no significant difference in moral responsibility attributions between Zygote and control cases. Thus, it seems that the very starting point of the Zygote argument - that we have the intuition that agents in Zygote cases are not morally responsible for their actions - is simply a non-starter.

Of course, defenders of the Zygote argument could simply argue that folk intuitions about the Zygote cases are irrelevant, and that what counts is intuitions from trained, impartial philosophers with a good knowledge of the debate and the ability to understand the subtleties of such cases. This is precisely Mele's contention:

> Suppose an intuition check were to be run on premise (1). Would incompatibilists uniformly deem 1 true whereas compatibilists uniformly deem (1) false? Some philosophers and psychologists run

> controlled intuition checks on untutored subjects. I myself am doubtful about the significance of the judgments such subjects make about complicated theoretical matters. A more suitable audience for the question about premise (1) of the zygote argument might be people who have thought long and hard about freedom and moral responsibility and are agnostic about compatibilism. I call them *reflective agnostics*. (Mele, 2008: 280-281)

But we also have grounds for a *soft-line* response to the argument: indeed, participants did not consider agents in Zygote cases to be identical in all relevant respects to agents in control cases. If we focus on Bad Outcome cases (the ones for which we found an effect of manipulation), we can see that participants in Studies 1 and 2 were less likely to see agents in Zygote cases as being acting from their deep self, and more likely to see their mental states as 'bypassed'. Moreover, these differences seemed to play a role in participants' judgments: the effect of manipulation was mediated by Deep Self considerations, but also by participants' Bypass ratings (as I observed in additional, *post-hoc* analyses). Finally, we saw that the effect of manipulation was stronger for Bad Outcomes, compared to Good outcomes - exactly as I predicted based on the Deep Self Provenance model. Thus, even if participants' attributions of moral responsibility were affected by manipulation, this effect can be explained away by considerations available to compatibilists.

Once again, defenders of the Zygote argument could object that the relevant intuitions, for example those of Mele's *reflective agnostics* are not likely to be driven by the same considerations as those of non-specialists. But is it really the case? When discussing the intuitions of *reflective agnostics* (among whose he counts himself), Mele makes the following prediction:

> Thus far, I have not said what the event is that Ernie was built to produce (in the original story), event E. Different specifications of it may affect the strength of intuitions about the story. Suppose that E is the death of Ernie's aunt and that Ernie poisoned her in order to inherit her money so that he could get himself out of serious financial trouble. Some reflective agnostics will feel pulled toward the judgment that Ernie is not blameworthy for the killing because Diana assembled his zygote as she did to ensure that he would do precisely that and because her creative activity did ensure that he would do that. Furthermore, because they judge that if Ernie had been morally responsible for the killing or had killed his aunt freely, he would have been blameworthy for the killing, they feel pulled toward the judgment that the action was not free and not one for which Ernie is morally responsible. These same agnostics might have a different attitude if E were a homeless shelter's receiving a $200 donation and A were Ernie's donating that money. Other reflective agnostics may be more powerfully moved by Ernie's bad will in the killing scenario than by the details of his creation and be pulled toward the judgment that he is blameworthy and morally responsible for the killing and freely kills his aunt. (Mele, 2008: 283).

So, Mele predicts that reflective agnostics' intuitions about Zygote cases will depend on the kind of action performed by the agent. And, in this paragraph, he clearly suggests that reflective agnostic are more likely to think that some reflective agnostics will be more likely to consider the agent in Zygote cases to be morally responsible when the outcome is good (e.g. donating to a homeless shelter) than when the outcome is bad (e.g. killing one's aunt). But, this is exactly what I observed in Study 1. Thus, it does not seem that these hypothetical reflective agnostics' intuitions are much more different than laypeople's intuitions. So why think they have different psychological underpinnings?

Moreover, Mele does not offer any plausible mechanism that would explain why these reflective agnostics would be more willing to attribute moral responsibility when the agent is giving money compared to when they kill their aunt. To my knowledge, the only available explanation we have at the moment is the Deep Self Provenance hypothesis, which predicted these very same results. As such, it seems legitimate to conclude (at least temporarily) that the intuitions of Mele's reflective agnostics are also driven by (implicit) considerations about whether agents in Zygote cases act on the basis of their Deep Self, which allows to reject Mele's claim that we have the intuition that agents in Zygote cases lack moral responsibility even when they fulfill all (non-historical) compatibilist requirements. Rather, it is probably because we perceive (consciously or not) as being prevented to act on their own deeply-held desires and values.

Finally, the data presented in this paper give us a third reason to be skeptical of the Zygote argument: indeed, the evidence suggests that the intuition that agents in Zygote cases are not morally responsible for their actions has nothing particular to do with determinism. In Study 1, the Indeterministic case was not significantly different from the Manipulation case, but was significantly different from the Control case. In Study 2, there was no significant difference between the Manipulation and Indeterministic case. Thus, it seems that, whatever drives participants' intuitions about Zygote cases, it has nothing to do with determinism. Thus, determinism cannot be counted as the best explanation why agents in Zygote cases are not morally responsible, and the explanatory version of the Zygote argument is no more successful than the direct one.

Of course, all these counter-arguments to the Zygote argument might once again be dismissed on the ground that folk intuitions do not matter, and only experts' intuitions are conducive to truth. However, I would resist this defense on two grounds. The first ground is meta-philosophical: doing so seems to assume that there is some kind of metaphysical (or rather, moral) truth about moral responsibility only expert philosophers would have access to (or to which expert philosophers would have better access). I must say that I have always failed to see what would warrant such a view of the philosophical enterprise: our notion of moral responsibility is not a technical one, but rather comes from attitudes and practices that preexist the philosophical enterprise. The second ground is dialectic: as an expert in the field myself, I have never shared the intuition at the basis of the Zygote argument, and I am not the only expert in this case. Knowing that common sense is my side, why should I even worry with the Zygote argument to begin with?

Acknowledgments

Florian Cova's work on this chapter was supported by an Eccellenza Grant of the Swiss National Science Foundation ("Eudaimonic emotions and the (meta-)philosophy of well-being"). I would

also like to thank Yoko Takahashi and AmaLee for their emotional support and for inspiring the title of this chapter.

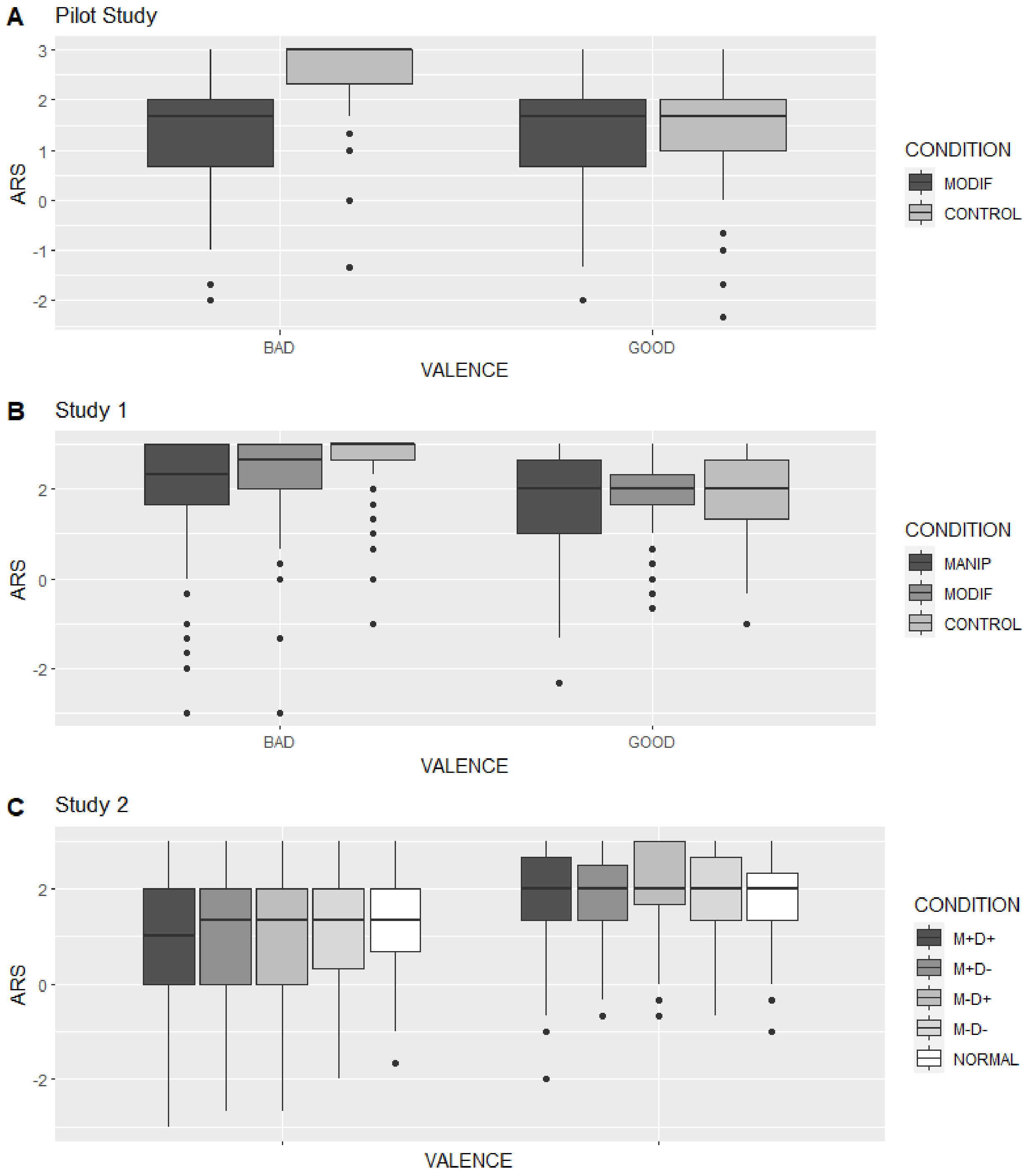

Figure 3. Boxplots of Aggregate Responsibility Scores (ARS) in function of CONDITION and OUTCOME VALENCE for all three studies.

References

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

Björnsson, G., & Pereboom, D. (2014). Free will skepticism and bypassing. In W. Sinnott-Armstrong (Ed.), *Moral psychology*, (Vol. 4, pp. 27–35). Cambridge, MA: MIT Press

Cova, F. (2011). *L'Architecture de la cognition morale*. PhD thesis. EHESS.

Cova, F. (forthcoming). Are folk libertarian compatibilists? In J. Campbell, K. M. Mickelson and V. A. White (Eds.), *Blackwell Companion to Free Will*. Blackwell.

Cova, F. & Boudesseul, J. (2021). "That feels deep!": Feelings of being moved play a role in perceptions of depth and profundity (feat. the Geneva Sentimentality Scale). Unpublished manuscript, University of Geneva.

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

Kane, R. (1985). *Free Will and Values*. Alabany, NY: State University of New York Press

Kearns, S. (2012). Aborting the zygote argument. *Philosophical Studies*, *160*(3), 379-389.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. OUP Oxford.

Mele, A. R. (2005). A critique of Pereboom's' four-case argument for incompatibilism. *Analysis*, *65*(1), 75-80.

Mele, A. (2006). *Free will and Luck*. New York: Oxford University Press.

Mele, A. R. (2008). Manipulation, compatibilism, and moral responsibility. *The Journal of Ethics*, *12*(3-4), 263-286.

Mele, A. R. (2019). *Manipulated agents: A window to moral responsibility*. Oxford University Press.

Mickelson, K. (2015). The Zygote Argument is invalid: Now what? *Philosophical Studies*, *172*(11), 2911-2929.

Mickelson, K. (2017) The Manipulation Argument. In: M. Griffith, K. Timpe, and N. Levy (Eds.), *The Routledge Companion to Free Will*, pp. 166–78. New York: Routledge.

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.

Pereboom, D. (1995). Determinism al dente. *Noûs*, *29*(1), 21-45.

Sripada, C. S. (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*(2), 159-176.

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

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, *12*(4), 551-560.

Turri, J. (2017a). Compatibilism can be natural. *Consciousness and Cognition*, *51*, 68-81.

Turri, J. (2017b). Compatibilism and incompatibilism in social cognition. *Cognitive Science*, *41*, 403-424.