Charles University

Faculty of Arts

Department of Philosophy and Religious Studies

# BACHELOR THESIS

# BAKALÁŘSKÁ PRÁCE

Hana Kalivodová

## Manipulation and moral responsibility

## Manipulace a morální odpovědnost

Praha 2021 Supervisor (Vedoucí práce): doc. Mgr. Tomáš Marvan, Ph.D.

Prohlášení:

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného či stejného titulu.

Declaration:

I declare that the following BA thesis is my own work for which I used only the sources and literature mentioned, and that this thesis has not been used in the course of other university studies or in order to acquire the same or another type of diploma.

Prague, 6. August 2021 ………………………

Hana Kalivodová

I would like to express my gratitude to my supervisor, doc. Mgr. Tomáš Marvan, Ph.D., for his willingness to supervise the thesis, his help and patience and the courage to read the first drafts of the chapters.

I also would like to thank all the nice people who took time to point out all the incopherensibilities and helped me significantly to increase readability of the thesis and decrease the number of 5-row sentences. Namely Anton Khirnov, Patrick Edwards, Barbora Hanzalová, Filip Štědronský and Karel Murgaš.

Abstract:

In recent philosophy debates, conflicting views persist on the influence of manipulation on moral responsibility of individuals. One side sees manipulation as not different from any other deterministic environmental influence on the agent, others advocate for historicism — the idea of deviant causal route — if manipulation is present in the agent's history. Historicist account of moral responsibility is based on disruption of natural development, but this may not be trivial to detect in cases of covert non-constraining control and social conditioning. In addition, opinions on the resulting responsibility of the agent are influenced by intuitions on her identity, locus of control and nature of motivating reasons. The aim of this thesis is to map the recent debate and identify the factors playing the role while searching for the borderline of responsibility of the manipulated agent.

Keywords: manipulation, moral responsibility, causation, control, historicism

Abstrakt:

V současné filozofii probíhá debata o vlivu manipulace na morální odpovědnost jedince. Jedni poukazují na to, že manipulace není odlišná od jakéhokoliv jiného deterministického působení prostředí na agenta. Jiní přicházejí s historicismem, ideou, že odpovědnost je citlivá na výskyt manipulace v kauzálním řetězci předcházejícím činu. Historicismus stojící na narušení přirozeného vývoje může být ale těžko obhajitelný v případech skrytého společenského podmiňování. Názory na konečnou odpovědnost agenta navíc ovlivňují intuice o jeho identitě, kontrole a původu důvodů, jež ho motivují k činu. Předmětem práce je současnou debatu zmapovat a pojmenovat faktory vstupující do hry při hledání hranice odpovědnosti manipulovaného agenta.

Klíčová slova: manipulace, morální odpovědnost, kauzalita, kontrola, historicismus

# Contents

# 1. Moral responsibility of manipulated agents

## 1.1 Moral responsibility

### 1.1.1 What is moral responsibility

What does it mean to hold someone morally responsible? As opposed to legal responsibility, the moral one has only unwritten rules, which may cause differences in opinions about what is relevant when ascribing it. There are no canonical ways nor authoritative judges. As opposed to causal responsibility, the moral one includes certain normative force. It seems to be at least a prerequisite for further moral judgements about the agent. Beyond that, the definitions may vary, but it seems to be the core of our common moral practices.

Are there any reasons to participate in holding others morally responsible or is the whole concept an illusionary rationalization of unsubstantiated human custom? I do not want to break down the Chesterton's Fence[1] here. I suppose there is a system created by someone for some purpose and while we may have data for some minor changes, taking the whole practice away can cause serious problems in unforeseen places. There are certainly crucial society-sustaining powers in the practice and there are also significant critiques of the concept, especially of the practice associated with it - praising and blaming[2]. While there may be better ways to react to particular actions of an agent than praise and blame, keeping each other morally responsible as a shared conceptual idea can have preventive power on its own. We can also understand moral responsibility as a connection between agents and their actions and secondarily the consequences of their actions, or more specifically

[1] Chesterton’s Fence is a concept inspired by G.K. Chesterton’s story about a person who desired to intervene and destroy a fence they did not see the use of. The point of the heuristic is not to remove any system we have, unless we understand why it was built in the first place.

[2] For example, Bruce Waller’s book

between agents and their psychological states (intentions) regarding their action without any relation to some consequent social reaction (such as blame). Nevertheless, some philosophers understand moral responsibility simply as blameworthiness[3], which would be discussed in detail in chapter 2.2.

Moral responsibility is not attributed to agents by default. We can see it in contrast with causal responsibility. Imagine a drunk relative tripped you on purpose at a family party and you fell and broke the leg of a baby who was crawling on the carpet. You are only causally responsible in this case, because you did not choose an action even remotely similar to the outcome and you could hardly have prevented any of it happening. It does not apply to your will either directly or in the sense of the omission. Mele suggests that feeling terrible in this scenario, given what happened, is normal, but it should not be mixed with feeling guilt[4].

### 1.1.2 The origin of the debate

The roots of the debate lie in the problem of free will which has been here with us ever since the origins of man. It started with old religious thinkers and the idea of predestination and similar ideas leading to fatalism after opening a question: "If God controls our lives and fates, why would we even try?" and transformed into a similar problem with scientific causal determinism.

Philosophers gradually revealed a number of different things which free will could mean. As a result, they defined that the kind of free will particularly endangered by determinism is the *free will needed for moral responsibility*. Without free will and free action, all our moral practice will seem meaningless.

In the 20th century, the following formulation of the problem of free will became

[3] Alfred Mele, *Manipulated Agents* (New York: Oxford University Press, 2019), p. 3

[4] Mele, *Manipulated agents*, p. 23

established:

- Causal determinism means that if we know all physical laws and the initial state of the world, we know all the following states, because those are completely defined by the physical laws and previous states of the world.
- Free will means that we are the source of our action, our action is not defined by things beyond us.
- Thus determinism rules out free will. If we live in a deterministic world, we cannot have free will.

To save the room for free action, one can employ several strategies. The first view, called *libertarianism*, denies that determinism exists. With the reference to quantum physics, they mostly appeal to indeterministic phenomena, at least in the brain. The second view, called *compatibilism*, accepts determinism and tries to redefine free will in terms compatible with it. Mostly, they appeal to some kind of authenticity and claim that indeterminism is just a kind of randomness and indeterministic free action is not truly ours[5]. The last view is held by those who do not want to save free action. It is called *hard determinism* and it states that since free will is not compatible with determinism, it simply doesn't exist. They may call it illusory and still find reasons for societal practice of holding each other morally responsible or they may criticize even the practice.

Despite the specification that we are concerned with *free will needed for moral responsibility* only, the term *free will* remains ambiguous. Since the growing number of philosophers did not feel the need to fight for it in its old, as Daniel Dennet would say, magical[6], form, in the recent papers, the problem is discussed increasingly as

[5] Daniel Dennett, *Freedom Evolves* (New York: Viking Books, 2003)

[6] In his book Freedom Evolves, Dennett uses an analogy of an elephant Dumbo who flies thanks to waving his giant ears, but thinks that a magical feather he holds does it. Dennett thinks that we are using freedom in the same superstitious way when we believe that it makes us an indeterministic source of our action.

just moral responsibility. After a period of establishing so-called classical compatibilist and classical libertarian accounts, a lot of critique followed and new, less radical positions such as Semicompatibilism or Modest Libertarianism emerged. The original three views branched out and now it seems that almost every philosopher has his own view with his original approach. The focus on the questions of determinism versus indeterminism and compatibilism versus incompatibilism dichotomies fades slowly out and the emphasis is now more on different conditions or capacities which make an agent a morally responsible being.

### 1.1.3 Ruling out alternative possibilities

The new era started with Harry Frankfurt's compatibilist defense of the claim that moral responsibility does not require the *possibility to do otherwise,* a standard presupposition treated as given by both, compatibilists and incompatibilists. His 1969 examples are still discussed lively and referred to as *Frankfurt cases*. The case features an agent, Jones, who is considering a certain action. Another agent, Black, wants him to do that and is prepared to intervene, if he decides otherwise. Eventually, Jones decides to perform the action Black wants him to do on his own. His behavior was voluntary and it does not seem reasonable to excuse him from moral responsibility, yet he "could not have done otherwise", because Black has the powers to ensure he will have his way.

The conclusion is that it does not matter if the agent is causally determined. What matters is if they acted out for their own reasons. Other philosophers followed with their own ways to reject the *possibility to do otherwise* or as it was called by Frankfurt: *Principle of alternative possibilities*[7].

[7] Principle of Alternative Possibilities (PAP) defines as "a person is morally responsible for what she has done only if she could have done otherwise.", https://plato.stanford.edu/entries/alternative-possibilities/

Semicompatibilists such as Fischer claim that moral responsibility is possible even though determinism is incompatible with the ability to do otherwise[8]. Fischer and Ravizza are distinguishing *guidance control* and *regulative control*. While regulative control means both the power to perform an act and the power to freely do something else instead, guidance control does not require access to alternatives. We can imagine an agent driving home and guiding themselves to turn right regardless of whether it was open to them to turn left[9]. For Fischer and Ravizza, only the guidance control is necessary for moral responsibility.

In any case, we can imagine having alternative possibilities in different degrees of realism, or lack there of. From being open to anything, so I can toss a coin and act accordingly, to being physically tied up somewhere, but with the option to bite my little finger off and escape if I really want to. Philosophers gradually came to the conclusion that the difference between free and unfree agency lies in how much we act on the basis of our own psychological structures (such as beliefs or desires) compared to what others want from us or what the situation demands.

One of the greatest contributions to the discussion was made by Alfred Mele. He was the first one who refused to support either determinism or indeterminism and instead argued for both sides and tried to find accounts of free will and moral responsibility for both libertarians and compatibilists. In his book *Autonomous Agents: From Self-Control to Autonomy*, he supports Daniel Dennett's idea that if there is some indeterminism along the way to our decision, it should be in the early phase of generating the options, not immediately before the decision itself as suggested by libertarians.

[8] Fischer, John. 1994. The Metaphysics of Free Will. Cambridge, MA: Blackwell. 1994, p. 180

[9] John M. Fischer a Mark Ravizza, Responsibility and Control: A Theory of Moral Responsibility ( Cambridge: Cambridge University Press, 1998), p. 33

He claims that we can have a regular free will defined by negative conditions, in the sense that we were not forced to do so, without having the kind of free will suggested by the *Principle of Alternative Possibilities*.
He described that the idea of *alternative possibilities*, which is something libertarians really cling to, means for them that we can do otherwise in the possible world with the same past and laws of nature up to the moment of the decision. But this is not needed, we can contend that a broader range of worlds is admissible for tests of the relevant kinds of ability[10]. We can have the freedom needed for moral responsibility only in a sense that we could have done otherwise in *similar possible worlds*.

Another way to explain the sense in which we have the freedom needed for moral responsibility even under determinism is by comparison with a compulsive person. Mele writes that the dif a compulsive hand washer, is that he washed his hands freely and his action had some moral significance, although both could have been causally determined to do so[11]. I would add that the moral significance is there because there were some reasons behind Mele's decision to wash his hands.

### 1.1.4 The current debate and Vargas' point of view

Thanks to the *Frankfurt cases*, the attention turned to reasons for an action. Accounts which have the ability to respond to moral reasons in their center are called *reasons-responsive*. It seems that the majority of the theories at least count on some reasons-responsiveness. It could be characterized as "capacities for being appropriately sensitive to the rational considerations that bear on their actions[12]". Besides this characteristic, moral responsibility is seen increasingly as a *social practice*. It was not always the case. The idea that agents exercise some kind of

[10] Mele, *Manipulated agents*, p. 18
[11] Mele, *Manipulated agents*, p. 28
[12] https://plato.stanford.edu/entries/moral-responsibility/#ReasRespView

metaphysical control, and based on that, they truly deserve praise or blame was quite common[13]. These accounts were called merit-based views and it seems that they are only slowly in retreat thanks to the critics such as Pereboom or Waller. The alternative approaches claim that there is nothing which makes an agent truly deserve some reaction, but it is appropriate to react if it would likely lead to changes in agent's behavior. It assumes that there is a human who is able to learn and modify his behavior in the future thanks to the reactions, and maybe some other humans who are looking and learning which behavior gets praise or blame. The point is to give the agents incentives to act on, encouraging them in certain choices to secure positive behavioral outcomes in the future[14].

Manuel Vargas summarizes it as follows: "We have all we need if we hold that blame is beneficial to the extent to which agents are moved to improve their behavior."

His book *Building Better Beings* is one of the recent efforts to build a theory of moral responsibility on the foundations of *reason-responsiveness*. His innovative theory is standing on revisionism of the notion of free will inspired by Daniel Dennett. He describes that our responsibility practices are justified by appealing to their suitability for fostering moral agency and the acquisition of capacities required for such agency and calls it the *agency cultivation model*[15].

Vargas' approach is strong in the social dimension of moral responsibility. He claims that it is dependent on the features of the agent as well as on the features of the environment. Yet, the environmental features he is discussing are mostly about the social environment and he does not provide many situational clarifications. I will also argue further in the text that his reactions to manipulation cases depend on

[13] Gary Watson, "Free Action and Free Will", *Mind* CXVI (1987)
[14] https://plato.stanford.edu/entries/moral-responsibility/
[15] Manuel Vargas, Building Better Beings: A Theory of Moral Responsibility (Oxford: Oxford University Press, 2013), p. 3

*reason-responsiveness* way too much. In spite of that, Vargas did a good job in mapping all the recent issues in moral responsibility theory. He also provided valuable comments on various manipulation cases, therefore *Building Better Beings* will be the primary source of the thesis.

After the *principle of alternative possibilities* lost relevance, new interpretations of free agency emerged. In the search for a new solid basis on which we can explain moral responsibility, multiple new definitions of conditions for moral responsibility were stated and this started the era of manipulation arguments. In one of the larger branches of the discussion, so-called manipulation arguments gradually prevailed. They appeal to our intuitions and show that the definitions of a morally responsible agent established in various, mostly compatibilist, accounts are not sufficient.

## 1.2. On manipulation

Manipulation could be perceived as something precluding moral responsibility but there is no consensus about it. It is more subtle than coercion and often seen in advertising, politics and in both professional and intimate relationships. If physical restraint and threats are already seen as something elevating moral responsibility from the victim, but the villains' strategies are changing and employing manipulative strategies on scale, we should consider to specify the borders in a different way than via the concept of free will. The problem is social and mass media influence which can be abused in noncoercive ways, as shown by the recent research in psychology, behavioral economics, and cognitive science. Thanks to its range, it can have even worse consequences than forcing an individual physically.
Allen Wood says that advertising corrupts the root of rational communication. It manipulates people into acting on impulses and in their most immediate and self-interested preferences, discounting both our own long-term interests and the

interests of others. It leads to inequalities in wealth and power[16]. The long-term harm is not only in the debt traps and destroyed families as a final destination for those who cannot see through business models of various corporations. Political advertising also embodies serious risk of radicalization and lives destroyed not only to the consumers of the advertising.

A recent example is a case of Mr Balda, a retired citizen of Czechia who lived in fear caused by anti-islamic political advertising, and decided to derail a train and made it appear that islamists did it to make people believe in the threat as well[17]. How was he manipulated? By certain speech, suggestive pictures, chain emails? Manipulation is not easy to specify, but while it is difficult to find some common features of it, it doesn't mean that there is nothing.

Manipulation entails some kind of moral failure itself, manipulating person acts on dishonesty instead of explaining and grounding the desired outcome we need others for. There is a discussion going on if manipulative acts are wrong by definition or if it just implies some moral reason to refrain from. Being manipulative may be distinct from its moral status, when we for example admirably manipulate someone into sitting quietly[18]. We can have the same discussion about white lies, but I want to focus on the moral responsibility of the manipulated agents, regardless of if manipulation is itself wrong as a tool or just thanks to its goals. If the manipulator is responsible is a separate question.

### 1.2.1 Global manipulation

[16] Allen Wood in Christian Coons-- Michael Weber, *Manipulation: Theory and Practice* (Oxford: Oxford University Press, 2014)

[17] https://www.bbc.com/news/world-europe-46862508

[18] Coons, *Manipulation: Theory and Practice*

In the real world, we know we may be manipulated by others to do things we would not have done, but their arguments and other persuasions changed our mind. By current definitions, this does not rob us of free will or the ability to act freely in the way required for moral responsibility. We could have resisted the sales pitch or subtle manipulative press and might blame ourselves later for not doing so[19]. To emphasize that the agent (victim) was deprived of free will, manipulation arguments introduce *global manipulation*. Those imaginative scenarios often include covert intervention implementing radical reprogramming of agent's beliefs, desires, and other mental states via neurological engineering[20].

### 1.2.2 Manipulation cases and its kinds

The idea of manipulation arguments is to assemble a pair[21] of cases where agents did the same action. In the first, the agent freely decided to do so, in the second the agent was manipulated in a way that we would not intuitively see him morally responsible. If the tested condition for moral responsibility is a valid one, it should not be present in the latter case and the agent should be rendered non-responsible. However, the author of the argument claims that both agents fulfill the condition of the theory for being morally responsible and there is no relevant difference between them according to the theory. This means the theory, or at least its condition for moral responsibility, is wrong.

Christopher E. Franklin distinguished three types of global manipulation arguments. (1) God-like manipulation argument where the agent is designed by a powerful being which can predict all his actions and design them in a way which will lead to desired events.

[19] https://plato.stanford.edu/entries/incompatibilism-arguments/#ManiArgu
[20] https://plato.stanford.edu/entries/ethics-manipulation/#OrdiVersGlobMani
[21] or a set

(2) Brainwash-like manipulation argument introducing evil neurosurgeons or psychologists which covertly manipulate agent’s psychological states.
(3) Natural causes manipulation argument which employs natural but extraordinary causes such as brain tumor, which can influence agent’s reason-responsiveness[22].

The characteristic of the cases which have all the types in common is that the manipulation is covert, preventing the agent from reflecting on their situation. If we lack critical information about the situation we are in, we can follow the steps we otherwise would not follow.
For example, if we often see in the media information about oceanian terrorism and illigal immigration in our state, there is a reason to think that we are in danger. But if we know additional information that it is a strategy of politicians who own the media to evoke this feeling in us, the probabilities will change and we can change our behavior completely as Bayesian reasoning under uncertainty tells us.

Another common feature is that there is another intentional agent and there is an asymmetry in powers between the manipulator and the agent which is abused. It could be misuse of information, knowledge, power or just the situation.

The role of the manipulation cases is not only to test and reject theories and their condition for moral responsibility. We can take it the other way around and test our intuitions about the cases. Regardless of what our strategy is, counterexamples can also serve us as an important indicator of the knowledge gaps in how we understand human decision-making. In addition, the reactions to manipulation cases are a valuable peek into the thinking of various philosophers. It shows us which aspects of the situation play a role in their accounts outside of the circumstances they modeled for themselves and what is important for them.

[22] Christopher E. Franklin, “Plausibility, Manipulation, and Fischer and Ravizza”, The Southern Journal of Philosophy 44 (2006), p. 173-192

Because manipulation is a special case where the agent themselves is seriously mistreated, it forces us to think about them as not so idealized agent able to process all the information and rationally evaluate their desires. If too much mistreatment is present, we should not ask if the agent is responsible but who from the influencing agents in the causal chain is responsible for the result.

### 1.2.3 The Zygote argument and the perspective of a bounded being

I mentioned that we are bounded beings. By that I mean that we have limited resources and limited points of view. We have to reach a decision in a limited time with limited information, attention and cognitive capacity. This applies to us as an agent and also as a moral judge. One kind of manipulation argument introduces God-like powers. It is, for example, Mele's Zygote argument:

A Goddess Diana placed a zygote in Mary, because she knew all the causal routes in the future including that Ernie will be born and at the age of 30, he will perform an act, which Diana desired.

Mele suggests that in comparison with Bernie, who went through the same causal routes as Ernie and did the same thing at age of 30 but was born without Diana's intention and intervention, Ernie is not morally responsible for his action. The problem I have with this case is that we are judging it from a god-like perspective. In a normal world, we will never know what Diana did. If it really happened, we will make an imperfect judgement and perhaps be unfair to Ernie. But it is ok, we are bounded beings, the common practice is to judge agents morally responsible from grounds we have, not from ideal grounds we cannot even imagine being real. Of course if we get to know that Diana did it, we can judge Ernie not morally responsible. However, there is a chance that in a world, where we can overhear goddess Diana’s bragging about how she put the zygote in Mary, some different moral rules will exist, since you have to expect people being toyed by Gods. There is one more reason to think that we should still judge Ernie for his deed: the forward-looking effects. Maybe Ernie performed the act Diana wanted now, but we can prevent it happening in the future.

Note that Mele’s definition of moral responsibility is “being blameworthy”. Is this the kind of moral responsibility worth wanting? Ernie could be responsible in a broader sense: If he now has a structure allowing him to perform the act Diana wanted and the structure could be reformed, it makes sense to hold him morally responsible at least in a sense that he is the one who should be shown how to do things properly. In addition, people who are judging characters more than separate acts would be inclined to judge Ernie nevertheless.

The original goal of the Zygote argument was to show that Ernie fails to be a source of his action thanks to the manipulation. It is also meant to show that there is no difference relevant to many compatibilist accounts between a person with a manipulated zygote and any person in a deterministic universe. That is because we are defined by the previous stages of the universe in both cases, and don’t have free will in the sense of *alternative possibilities* or ultimate control over the situation. Ernie is described as a person who fulfills all the conditions for moral responsibility a Frankfurt-like compatibilist has: he is mentaly healthy, self-controlled and definitely not compelled or coerced to perform the act. But still, he seems to be failing to be a proper source of his action and Mele thinks compatibilists have to bite the bullet by saying he is nevertheless responsible[23]. I would like to take two things from this: (1) social practice is always practiced from a human point of view, we have to accept that we do not have all the information about the agent and we are sometimes judging them even though they were determined or

[23] Alfred R. Mele, “Manipulation, Moral Responsibility, and Bullet Biting”, *The Journal of Ethics* 17 (2013), p. 167-184

non-deterministically caused to do that and (2) that it is all right because this is a core of social practice and by giving them feedback, we are also giving them the incentive to rebel against the causal factors and consider other options in the future. Even without a goddess Diana, we could have been defined by a more sophisticated person at the moment and moral social practices exist here to make people stop, think and evaluate any negative influences on their life which could be reverted by reflection. Instead of drawing the conclusion to one side by saying that Ernie is also morally responsible or to another by saying that Bernie is also not responsible, we can focus on the difference[24] in the freedom status of the two agents and look for ways to prevent the situation so that more people are not manipulated by Diana.

### 1.2.3 Definition of the playing field

**Ruling out determinism and indeterminism discord**

There are already various examples showing that modified manipulation cases could be a bullet-biting problem also for libertarian accounts of free will. They are discussed in papers by Haji and Cyupers[25] or Mele's student Taylor Cyr[26]. For illustration, here is a variant of the Zygote Argument mentioned by Cyr and originally constructed by Stephen Kearns[27] which works in an indeterministic scenario:

> "Diana creates a zygote Z in Mary. She combines Z's atoms as she does because she wants the zygote to develop into an agent who performs a certain set of actions over the course of his entire life. From her knowledge of the

[24] And I think that the difference is not in the structure of the agent or it is certainly not the thing that would help us in the prevention program.

[25] Ishtiyaque Haji -- Stefaan E. Cuypers, "Libertarian  Free Will and CNC Manipulation", *Dialectica* 55 (2001), p. 221-238.

[26] Cyr

[27] Stephen Kearns,. "Aborting the Zygote Argument", *Philosophical Studies* 160 (2012), p. 379-89.

> state of the universe just prior to her creating Z and the laws of nature of her *indeterministic* universe, she deduces that a zygote with precisely Z's constitution located in Mary will develop into an ideally self-controlled agent, Ernie. As Ernie lives his life, there is a small chance every few seconds that Ernie is incapacitated due to the way Diana created his zygote. If Ernie is never so incapacitated, then he performs that set of actions that Diana has planned. As it happens, Ernie is never incapacitated and performs all those actions Diana has planned. (2012:385 , emphasis original)"

We can modify even the original *Frankfurt case,* where Black is watching over Jones and is prepared to intervene, to make it indeterministic. I believe it was done originally by Peter van Inwagen and the only difference is that Black has only a 98% chance that his intervention would be successful. I consider these arguments convincing and believe that we can discuss the issue of manipulation and moral responsibility without considering determinism or indeterminism. We can be agnostic about it just as Mele was from the beginning.

**Defining the goal**

In the next chapter, I want to analyze factors that play a role in judging manipulated agents responsible or not responsible and present which of the factors are important for philosophers. Particularly for Manuel Vargas, and Alfred Mele as an author of the manipulation cases who provided a powerful critique of the conditions for moral responsibility defined by various philosophers, namely Harry Frankfurt.

All of this is just to prepare the grounds for analyzing manipulation cases and showing through them that moral responsibility is a complex phenomenon which depends heavily on other concepts in our theories. Even concepts seemingly distant from the domain of ethics make us assume some features of responsibility and this may be revealed by borderline cases, such as manipulation ones. Some undeniable ambiguity is inherent to the concept of moral responsibility, but manipulation cases show us the limits which can at least in a negative way say something about it, or at least about philosopher's ideas about it.

There are two approaches to define moral responsibility: (1) by negative conditions (the same way as we excluded physical restraint and coercion) and (2) by positive structural conditions. I would like to show that manipulation could be a candidate for moving among the negative conditions.

# 2. A unique way to fail in doing good

## 2.1 Three accounts versus factors influencing our judgement

I mentioned several approaches to moral responsibility. The greatest detail will be given to Vargas' *Reasons account.* It will be compared mainly with Mele's historicist account and Frankfurt's *Real Self* account.
The goal is to describe what a function of blame is for them, what their understanding of control is and which other factors are influencing their final judgments of manipulated agents. There are many aspects of responsible agency we can focus on. It is even hard to discover which of the aspects are the same, but named differently by various philosophers. This chapter will start with Vargas' concept of blameworthiness, which contrasts with those of Mele and Frankfurt and judges manipulated agents from a completely different angle. I will go through several differences caused by whether we imagine moral responsibility as when the agent is accountable for something or something is attributed to them. Before I start discussing the particular factors, I would also suggest a neglected attribution problem, which can affect the further development of moral practice.

From the factors, I will discuss Frankfurt's definitions of control. Vargas is sceptical that there is a single control condition[28] or a cross-situationally stable mechanism constituting a unified capacity. His reasons-responsiveness is meant to be a cluster of more specific, ecologically limited capacities[29]. His account will be explained in detail in the next chapter. Mele has his own positive account on control (or maybe several ones he constructed to satisfy libertarian and compatibilist positions), but I will focus mainly on his original arguments to support a type of control which is usually neglected by philosophers - the character control.

[28] Vargas, Building Better Beings,p. 223
[29] Vargas, Building Better Beings, p. 205

Other two factors which I will mention in this chapter will be the overall capacity for moral responsibility and the concept of reasons.

### Conceptual factors

Several factors play a role in judging agents to be morally responsible for their actions. I would like to divide them into three broad categories: conceptual, external and structural. Conceptual factors are our assumptions or ideas about neighboring philosophical problems. Their details influence our concept of moral responsibility and our resulting responsibility judgement. I will start with those suggested by Manuel Vargas[30]. He wrote that "one's reaction to manipulation cases is partly structured by: (1) whether one operates with an *internalist* conception of reasons; (2) whether one imagines manipulation cases as *replacing* control structures; and (3) how one thinks about *personal identity*." But this is not a full list. Other factors could be (4) our conception of character, (5) whether we treat blame as forward- or backward-looking, which is connected with the question of whether we emphasize fairness or benefit, and (6) whether we focus on accountability or attributability. I would say that there are more hidden factors such as preferred moral theories which would play a role, but it is impossible to capture them all.

### External factors

External factors, on the other hand, are something which could be checked by investigators. They are everything which is already decisive in the investigation of crimes: use of force or threats, creating pressure, unauthorized intrusion into private property, concealment of mandatory information and similarly non-violent but still dangerous wrongdoings e.g. illegal selling methods. They include observable causal involvement and also observable consequences of the action. The

[30] Vargas, Building Better Beings,p. 278

consequences are somewhat controversial, because a significant part of moral approaches (e.g. kantian) is that one cannot blame an agent who caused something without ill will. These cases of omission or trying our moral luck are discussed in detail by Mele[31], but in ordinary moral practice, we just get more blame, if we were unlucky. The amount of blame for the actual consequences also depends on a role-specific expectations as recent studies show[32]. We blame managers more than ordinary workers. In general, everyone with a greater formal or informal role or from whom society expects more, gets more blame.

There are also historical external factors, something that happened long before the event and formed the agent's beliefs, character and the way they are interpreting their perceptions. I believe that it is to some extent the same thing as the conceptual structures we suppose in the agent. We can think the same of internal agential structure, a kind of learned reason-responsiveness, and a series of forming historical events that led to the current event. We can claim that the cause of an agent's selfish behavior is his psychological structure with selfish character or we can say that it is the continual behavior (series of forming events) of his pampering mother. In some cases the first one will be easier to check, but in the context of one-time events only the second could be used as evidence.

## Structural factors

Psychological structures are an especially problematic part. They are often assumed but rarely explained in sufficient detail. Some capacities of the agent are easily testable. We can see if a person we talk to is mentally ill. However, people could be impaired in many ways and that is why philosophers came up with definitions of

[31] Alfred Mele, Free Will and Luck (Oxford: Oxford University Press, 2006)

[32] Pascale Willemsen, Albert Newen and Kai Kaspar, "A new look at the attribution of moral responsibility: The underestimated relevance of social roles", *Philosophical Psychology* 31 (2018), p. 595-608

capacity for moral responsibility and several other conditions claimed to be constituted in actual or hypothetical[33] structures internal to the agent.
The capacity for moral responsibility is sometimes also discussed as autonomy[34]. There are two approaches to autonomy. First, the traditional one is individualistic and it focuses on the agent and its features only. That includes Frankfurt's account and basically all philosophy until the 20th century. The other one started to become popular in recent years and is called relational autonomy. Catriona Mackenzie defined it as follows: "persons are embodied, and socially, historically and culturally embedded, and their identities are constituted in relation to these factors in complex ways[35]". If we satisfy conditions for self-governance, we reach autonomy as both a capacity and a status concept. That means we are socially and politically recognized as having both authority over our choices and the power to act on that authority[36]. The conditions are usually defined as structural features of the agent according to the individualistic approach but they are in relation to the external world and situations in relational autonomy approach. I won't be using the term, but I wanted to point out that Mele's historicism and even Vargas' emphasis on the social dimension of moral responsibility are part of a bigger shift in current thinking. The right agential structures are considered the main internal condition for agents to be morally responsible, so how we imagine them is the biggest factor for the final judgement of responsibility of manipulated agents.

[33] In either way, we do not have a direct access into an agent's thoughts and self reports could be biased by default in their favor. I would consider structural factors being closer to conceptual ones, because the reconstruction of what is going on in an agent's head is, at least at the current level of science, a pure concept. Even the existence of a possible motive for a crime does not prove that the agent was really considering it.

[34] Alfred Mele,. "Autonomous Agents: From Self-Control to Autonomy", (New York: Oxford University Press, 1995)

[35] Catriona Mackenzie, "The Importance of Relational Autonomy and Capabilities for an Ethics of Vulnerability", in Catriona Mackenzie, Wendy Rogers & Susan Dodds (eds.), *Vulnerability: New Essays in Ethics and Feminist Philosophy* (New York: Oxford University Press, 2014), p. 33

[36] Katrina Hutchison, Catriona Mackenzie, Marina Oshana, *Social dimensions of moral responsibility* (New York: Oxford University Press, 2018), p. 61

## 2.2 Judging agents morally responsible

### 2.2.1 Moral responsibility versus blameworthiness

As I wrote in the first chapter, Mele and other philosophers treat moral responsibility the same way as blameworthiness. It is a valid definition if our goal is to analyze the state of the art. The advantage is that if we ask "Is this agent blameworthy?" instead of "Is this agent morally responsible?" it employs our intuitive grasp of social behavior and we feel that there is or is not a right moment to blame based on our experiences. However, it is just a proxy for judging agents morally responsible. I would define a morally responsible agent as an agent in the causal chain which could be reformed by reflection, the agent can learn based on our (or their own) judgement. If we investigate the crime scene, we look first for the suspects, and then decide what kind of intervention is suitable to prevent the event next time. Bruce Waller defines moral responsibility as what justifies blame and praise and adds that it is a condition for claiming what one deserves (e.g. giving praise). This is grounds for him and another important critical philosopher, Derk Pereboom, to reject the habits about moral judgements, or at least in Pereboom's case, the idea that agents truly deserve moral judgements. I would not say that they deserve the intervention whatever it is. Just deserts require a knowledge of an omniscient being, because who are we to judge what an agent truly deserves? Maybe the only intervention we all deserve is in an ideal world with an infinitely patient teacher who will take time to make us understand the consequences of everything. But as we are bounded beings, we are trying our best to give others the feedback and sustain our society. Blame and praise are just some basic but consensual tools for how to do it. We can find a better tool for giving others feedback, for example replacing the blame with explanation. It also depends on our limited understanding of the situation, so I would be careful in saying that someone deserves blame.

Luckily enough, the question of appropriate reaction to manipulated agents is here regardless of whether the agent *truly* deserves it.

A good illustration to show the differences could be children. They have the capacity to be morally responsible, but it is reasonable to assume that they are in a situation they never encountered. That would make them not-blameworthy, at least not in the character judging sense. We should make another intervention rather than blame: explain what situation they were in. We may also say in a davidsonian way that expressing that someone deserves praise or blame is no different from disquotational function of truth and we are basically just recommending a particular linguistic action, i.e. saying "you should blame him as well to create pressure". No matter how much I would like to see moral responsibility and blameworthiness be treated separately, it is not what Mele and Vargas do, so I also will treat the terms interchangeably.

Vargas claims that being responsible agents entails also considering ourselves a self-governed person. We have to be an element in its own right, conceiving ourselves as ones who can self-govern and then actually do govern ourselves in light of moral considerations[37]. For me, it implies that he recognizes that there is not a distinct switching point when we can say "from now on, you participate in the responsible agency system on your own". It is a mere ritual of adulthood. While it is necessary to claim ourselves generally responsible at some point, in some situations (especially cases including online communication, where our ordinary intuitions about whom to trust could be misleading), it is plausible to suppose that the agent is still learning.

More research on what, beyond the performing agent, caused an event can help us to target remedial efforts more efficiently and in a more fair way.

[37] Vargas, Building Better Beings, p. 230

### 2.1.2 Vargas' forward-looking blameworthiness

Although, in the rest of Vargas' book he treats blameworthiness in the same way as moral responsibility, in the chapter dedicated to *Blame and Desert*, he claims that in his view of a *Reasons account* "standards for blameworthiness (holding fixed satisfaction of the responsible agency standards) are more demanding than the standards for responsible agency.[38]" But he merely means that the fact that an agent is a responsible agent in general does not say anything about his blameworthiness in a specific situation.

He explains further that "In our present circumstances, we are already generally committed to the notion that some failures of knowledge are non-culpable, even if there is some sense in which the agent could have taken steps to make him- or herself aware of the relevant considerations."
He warns against too much pressure on epistemic condition, because if blame avoidance becomes extraordinarily difficult, we run the risk of discouraging widespread commitment to the practice of responsibility[39].

That means there is some room for agents to fail to respond to considerations that could have been anticipated and to be morally responsible in general without entailing blameworthiness in a particular situation. His demarcation criterion for blameworthiness would be, in the end, a defect in the agent's level of concern as required by morality, i.e. ill will. Vargas rejects Pereboom's notion of basic desert, but holds that a plausible account of desert is given by something called the social self-governance model of desert. In this model, the desert is based on prescriptive theory characterizing what sorts of practices make sense given the agency cultivation model and our current arrangement of our psychological dispositions,

[38] Vargas, Building Better Beings, p. 236
[39] Vargas, Building Better Beings, p. 237

inherited social practices, and the like[40]. This is way more flexible than imagining desert as a description of what we do or what we are. Even when Vargas identifies moral responsibility with blameworthiness, for me, it looks like this forward-looking account on moral responsibility leaves the possibility that something other than quality of will might serve as a better account of blameworthiness. This could be true at least for some group with a different arrangement of psychological dispositions, inherited norms etc. His social self-governance model of deserving blame is inspired by Christopher Bennett and has two aspects: (1) the expressive aspect which means that blame expresses our dissatisfaction with what the agent has done with his quality of will and (2) the communicative aspect of intending to tell our dissatisfaction or conviction that they rejected our shared norms of conduct to the agent. Vargas writes further in the book that communication could be costly and that it is not guaranteed at all that blame will be communicated. To sum this up, Vargas claims that blame is ordinarily deserved because, in creatures like us, blame plays a crucial role in our ability to self-regulate. Without blame, guilt cannot benefit the wrongdoer[41].

### 2.1.3 Responsible agency defined

In moral practice, moral judgements seem to be at least three different judgments in one: We judge the agent morally responsible, we judge the action morally wrong (or right), and we judge the agent responsible for the particular (wrong or right) action. All three judgments can have some conditions and philosophers have their own models of what plays a role and into which variables it breaks down. In general, all of them assume some kind of mental capacity as a condition for agents being morally responsible at all. It could be broken down into age, education, cognitive abilities etc. Dealing with actions is a task for moral theories and the conditions for an action being wrong usually include doing harm to others. What is important for

[40]Vargas, Building Better Beings
[41] Vargas, Building Better Beings, p. 263

my pursuit of manipulated agents is the question of judging an agent responsible for the specific action. It also has its own set of conditions usually including something as a control or free will condition and an epistemic knowledge about the matter either as a part of that control condition or as a separate condition. Those three judgements together will render the agent morally responsible for the action.

Mr Balda, our Czech terrorist out of fear of terrorism, is an example of mentally healthy human who probably had the overall mental capacity to understand the matter. He was certainly not under coercion and was not temporarilly psychologicaly incapacitated. The question is if he really had suitable epistemic knowledge to understand the magnitude and severity of the threat he was actually trying to fight against. In my understanding, he certainly lacked some kind of control. Instead of being served and informed suitably, he was probably a victim of targeted marketing, selected information and misinformative chain emails. I would consider him morally responsible but deserving regrets rather than blame.

However in Vargas' eyes, the output of those three judgments is the desert itself. As I wrote, it is not the desert in Pereboom's basic desert sense, he wants to understand it only as a "needed social reaction". He calls the schema *agentic moral desert* and thinks of it as construed of a three-part relation between a person, the desert basis (or the things in virtue of which one is deserving), and the thing deserved[42]. The difference between my picture and his is only that his "thing deserved" scales only from blame to praise, but I think that it could be whatever feedback or reaction could support moral agency. Vargas' thoughts about the topic are based on Strawson's and Strawson defines moral responsibility as "to be the proper object of the *reactive attitudes*, such as respect, praise, forgiveness, blame, indignation, and the like.[43]" Strawson is known for treating moral judgments as emotional reactions and this is also a significant factor in making Vargas judge manipulated agents

[42] Vargas, Building Better Beings, p. 250

[43] Peter Strawson, "Freedom and Resentment", *Proceedings of the British Academy* 48 (1962), p. 1-25

responsible. For me, the part “to be a proper object of” matters and I would like to imagine my emotional reaction to wrongdoing more as the detection of “a proper object” itself and then choose a more appropriate reaction than the emotional one.

If moral responsibility attribution is an interpersonal practice keeping society a good, stable, cooperative place, how exactly does it work? This is the answer of Manuel Vargas:

> “When we hold one another responsible, we participate in a system of practices, attitudes, and judgments that support a special kind of self-governance, one whereby we recognize and suitably respond to moral considerations.[44]”

It could be interpreted as a self-managing system where all of us are teachers of self-governance, judges and subjects of the judgments at the same time. However, it should be noted that Vargas’ statement that blame and praise make sense to us because of its effects in society[45] is a bit of a controversial forward-looking definition and others perceive it differently.

### 2.1.4 Attributability or accountability

Vargas’ approach goes beyond the familiar concept of accountability, usually treated as retrospective. It makes it possible for him to hold even manipulated agents accountable in most of the cases. Mele and others would not judge agents because it would be beneficial to do so.

Frankfurt also thinks that manipulated agents could be morally responsible, but from different reasons. He holds that an act is still attributable to an agent if they

[44] Vargas, Building Better Beings, p. 3

[45] Vargas, Building Better Beings, p. 5

were manipulated to issue it from their *real self*. This is because the aim of *Real Self* accounts is to determine whether behavior is attributable to an agent as opposed to the question whether to hold the agent accountable for it[46]. Gary Watson points out that since the focus is different, the important factors and even conditions for attributability and accountability may differ. *Real Self* accounts have in their centre reasons and desires and claim that only those issued from our real self are something we are morally responsible for. In this sense, accountability is broader, because its point of view is more interhuman and we are accountable even for events we chose with no strong opinion.

Both attributability and accountability are important for moral responsibility. While accountability solves the particular cases of wrongdoing, for attributability a particular bad deed is merely a data point telling us to check the character of the agent. If the agent was just trying, learning something or they apologize and it is clear that they in general know what is bad about the behavior, there is no need for intervention or other concern, because the character seems to not be too dangerous to live in society.

Frankfurt is afraid of more serious failure modes than something which could be obtained in a single wrongdoing. His concern is about character that could be constituted by convictions such as harming others being a privilege of the strong or that a certain groups of people do not deserve respect. If we reconstruct the question he is asking it could be said as "Does this agent need to be re-educated?". And it goes deeper than Vargas' system of quick feedback, where we merely ask if it would be beneficial to blame a particular agent for a particular action.

### 2.1.5 Attribution of moral responsibility and analogical data problem

[46] https://plato.stanford.edu/entries/moral-responsibility/#AttrVersAcco, last access: 1.8.2021

It could be also helpful to make it clearer what our attribution models on moral responsibility are. By attribution I mean assignment of the event to its causes, i.e. to the agents in the causal chain leading to it, this time from merely a statistical causal point of view. In reality, there are usually multiple agents in the causal chain leading to a bad outcome. While blaming the manufacturer and distributor of the knife seems unreasonable, I would not say so in the case of people who helped make us believe that it is ok to use the knife against other people. In Data Science, there are several attribution models helping us to choose what to count as a cause of an event (See figure X). The most common model was the one which looked only at the last item.

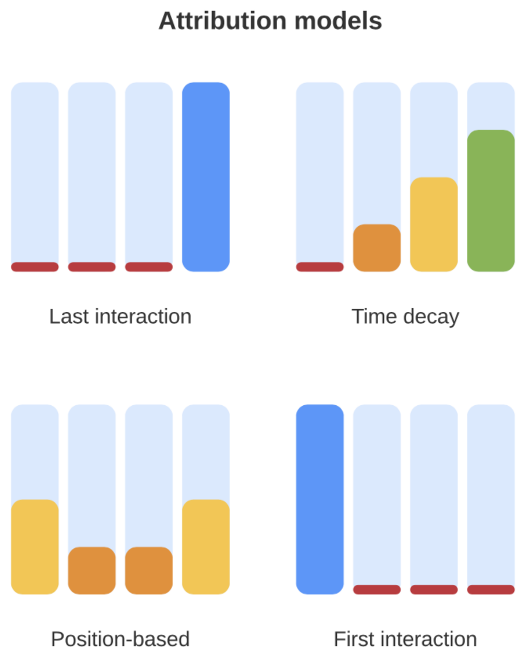

*Figure 1 - Attribution models*

Different models are suitable for different analyses and it is the same with moral responsibility. If we judge a child, there is an agreement that all the blame goes to the parent (first interaction gets all). If we judge an adult, they, usually, as the final agent in the chain who did the dirty work, get all the blame. But in more serious situations, the blame is distributed accordingly throughout the whole gang (and probably also the sentenced years). We tend to put all the pressure and blame on the performing agents themselves, because it is the easiest way for an average participant in a moral agency practice. We can see the performing agent without any investigation. Finding out who could have influenced the agent may be a futile effort which can result only in having conflicting claims. But in today's world when most of the communication happens online, we can have better data and prove what kind of communication and advertisements play the biggest role in influencing agents to commit offenses against other people. Then we can improve our moral practice. Not only by stopping the worst kind of influence, because we will know how many people will die on average if a person with half a million followers tweets something encouraging violent behavior. We can also move our attention and blameworthiness a bit from the manipulated agents at the end of the causal chain to the manipulators. If they play the game of numbers by spreading information to so many people that someone will act on it, society can do it as well. Manipulators are unscrupulously using social networks to spread dangerous informations to create in people the need for the products they offer (for example, for almost two-thirds of anti-vaccine content circulating on social media platforms is the responsibility of just twelve people, some of them owning businesses selling vitamins and alternative medicine[47]), but the actual deaths could be traced back to them and even if they are responsible for one thousandth of the particular bad outcome, it can make whole numbers again in the sum - we can count the number of people they killed indirectly but fully with such influence. And there are many other cases in politics and business[48].

The moral practice is often about judging the agent in front of us morally responsible or not. But social networks change the game, because we can give

[47] https://www.counterhate.com/disinformationdozen, last access: 6.8.2021

[48] The business models of betting applications and addictively designed (hazard) games are already known to be targeting the 4% of people in the population who have strong predisposition for addiction and will spend all their money in the app.

feedback to anyone. We do not have to ask only “Is the manipulated agent responsible?”, we can ask “Who are all the other agents supporting his decision and how much are they responsible?”. As both Frankfurt and Vargas noted, there can be more than one agent responsible for one bad outcome[49]. In addition, the same way the static models (which count only the easiest attribution) are being replaced in businesses by data-driven models, moral responsibility could also be determined based on the actual data and go beyond our intermediate reactions.

## 2.2 Conditions

I will now skip from attribution to the main factor or a complex of main factors playing a role in judging agents morally responsible: structural conditions. Maybe they are not mere factors, some philosophers consider them being the demarcation criterion itself. This is the case of Frankfurt. Vargas thinks, as I mentioned at the beginning of the chapter, that there is no way to find one simple criterion to separate morally responsible agents from those not responsible (or blameworthy). However, he tries to construct a system of structural conditions, a model of responsible agency, which is supposed to be flexible enough to work across situations. Mele, on the other hand, takes the view that structural conditions should be accompanied by external conditions which exclude manipulation from cases where moral responsibility applies. This negative condition (such as not having manipulation on out history) for moral responsibility will be discussed in the second part of the next chapter.

### 2.2.1 The epistemic and the control condition

[49] Vargas, Building Better Beings, p. 291; originally Frankfurt:”an agent can be fully responsible without being solely responsible” - Harry G. Frankfurt, *The Importance of What We Care About: Philosophical Essays* (Cambridge: Cambridge University Press, 1988), p. 25 n. 10.

The freedom condition, the most common structural condition for moral responsibility, was usually acknowledged as individually necessary and jointly sufficient for moral responsibility together with an epistemic condition. Epistemic condition is in different accounts called by different names: knowledge, cognitive, or mental condition. There is also no consensus if these two conditions are distinct because we may consider the epistemic condition being a part of the control condition. The epistemic condition is a general cognitive state which requires that we are aware of things we are doing on many levels, including foreseeing consequences of the action[50].

Vargas calls it self-directed agency a defines it as "in self-directed (but not yet responsible) agency are such things as beliefs, desires, means–end reasoning, the ability to formulate and execute action plans, and the presence of ordinary epistemic abilities, including a general capacity for some degree of foresight regarding the consequences of actions[51]"

He also includes other epistemic components into the control condition. Mele on the other hand wrote a critical article where he asked Fischer and Ravizza what epistemic requirements for being morally responsible for performing an action *A* are not also requirements for freely performing *A*[52]. Mele's answer is that there are no such self-standing epistemic conditions. Regardless of whether the epistemic condition is completely a part of a freedom/control condition or if we distinguish the epistemic condition for agency in general (like a nonmoral action of choosing an ice cream flavor) and the epistemic condition specific for moral agency (such as choosing a way to get some money), there are definitely some partial skills an agent can lack without missing the condition completely. It is also certain that some kinds of manipulation consist of manipulating the epistemic part alone. How it will affect

[50] https://plato.stanford.edu/entries/moral-responsibility-epistemic/, last access: 6.8.2021

[51] Vargas, Building Better Beings, p. 201

[52] Alfred Mele, "Moral responsibility for actions: epistemic and freedom conditions", *Philosophical Explorations*, Vol. 13, No. 2 (June 2010), p. 101–111, DOI: 10.1080/13869790903494556

the resulting responsibility will be discussed in the next chapter and I will now focus on the control condition itself.

Frankfurt presented several definitions appealing to a good integration into an agent's psychical condition, wholeheartedness or the unity of first- and second-order desires. All of them appeal to some kind of unity or authenticity in the agent and because of it they are classified as *Real Self* accounts. According to them, an agent is morally responsible if the behavior is attributable to their real (or deep) self.

The definitions are as follows:

1. Identification with the springs of the action

> "To the extent that a person identifies himself with the springs of his actions, he takes responsibility for those actions and acquires moral responsibility for them; moreover, the questions of how the actions and his identifications with their springs are caused are irrelevant to the questions of whether he performs the actions freely or is morally responsible for performing them.[53]"

2. Wholeheartedness

> "If someone does something because he wants to do it, and if he has no reservations about that desire but is wholeheartedly behind it, then—so far as his moral responsibility for doing it is concerned—it really does not matter how he got that way. One further requirement must be added: . . . the person's desires and attitudes have to be relatively well integrated into his general psychic condition. Otherwise they are not genuinely his . . . . As long as their interrelations imply that they are unequivocally attributable to him . .

[53] Harry Frankfurt. "The Importance of What We Care About", (Cambridge: Cambridge University Press. 1988.), p. 54

> . it makes no difference—so far as evaluating his moral responsibility is concerned—how he came to have them.[54]"

3. United first-order and second-order desires

> "[Humans] are able to form what I shall call "second-order desires" [...] Besides wanting and choosing and being moved to do this or that, men may also want to have (or not to have) certain desires and motives. They are capable of wanting to be different, in their preferences and purposes, from what they are. Many animals appear to have the capacity for what I shall call "first-order desires" [...] which are simply desires to do or not to do one thing or another. No animal other than man, however, appears to have the capacity for reflective self-evaluation that is manifested in the formation of second-order desires.[55]"

> "The unwilling addict has conflicting first-order desires: he wants to take the drug, and he also wants to refrain from taking it. In addition to these first-order desires, however, he has a volition of the second order. He is not neutral with regard to the conflict between his desire to take the drug and his desire to refrain from taking it. It is the latter desire, and not the former, that he wants to constitute his will. [...]
> The other addict is a wanton. His actions reflect the economy of his first-order desires, without this being concerned whether the desires that move him to act are desires by which he wants to be moved to act. If he encounters problems in obtaining the drug or in administering it to himself, his responses to his urges to take it may involve deliberation. But it never

[54] Harry Frankfurt. "Reply to John Martin Fischer." In S. Buss and L. Overton, eds. Contours of Agency. (Cambridge, MA: MIT Press. 2002.), 27–31. p. 27

[55] Harry Frankfurt, "Freedom of the Will and the Concept of a Person", *Journal of Philosophy 68* (1971), page 6-7

> occurs to him to consider whether he wants the relations among his desires to result in his having the will he has.[56]"

The last definition is called hierarchical. It is based on an appropriate reflection procedure which makes agent's motives, choices, or values authentically their own[57].

Both Vargas and Mele object to these conditions that our inner alignment or an active choice in accordance with our *real self* is not the only case where we are morally responsible. That is because they understand moral responsibility as accountability and we are accountable even for the actions that do not fully represent our inner character. For accountability, it does not matter if we had doubts and our real self is not (yet) a proper murderer. But it is ok, conditions for accountability and attributability may differ.

We can at least say that Frankfurt's conditions successfully exclude coerced agents from being morally responsible. You are definitely not wholeheartedly behind something you are coerced to do. However, there is still a broad spectrum of situations where an agent could be manipulated and still acting from his *real self.*

## 2.1.2 The capacity for moral responsibility

Holding agents morally responsible assumes some capacities or agential structures that enable the agent to even understand that he is held responsible and what it means.

Vargas criticizes Frankfurt's account in this regard. According to him, Frankfurt claims that there is *the capacity for reflective self-evaluation* required before one can have second-order desires.

Vargas sees no reason not to understand that (comparatively unusual) distinctive capacity as the locus of freedom. The reason for that is that the source of these

[56] ibid., page 12

[57] Hutchison et al.,Social dimensions of moral responsibility, p. 11

higher-order desires' relevance is exactly the fact that they are products of reflective self-evaluation. It would not support the distinction Frankfurt made if they were just a product of unmediated instinct[58]. But again, this seems to be a misunderstanding based on the confusion of accountability with attributability.

It seems clear that there is a difference between adult humans and animals or newborn children, because it makes no sense to hold the latter responsible. However, there is no consensus on where the borders of these capacities are. The *Real Self* accounts paid a lot of attention to the capacity. Susan Wolf presented a case of JoJo, who's *real self* is the product of a traumatic upbringing. She claims that while JoJo was raised by an evil dictator, he wants to be moved by torturing, imprisoning and other wrongdoings, but it still may be unfair to hold him responsible for his behavior, because he does not fulfill the moral competence condition. The issue is related to Mele's distinction between *sheddable and unsheddable values* and to the process of internalisation explained below.

At this point, we need to know that for Frankfurt the capacity for moral responsibility in the sense of accountability is a more basic condition, distinct from the control condition which he uses in a sense of attributability. This contrasts with Vargas for whom the model of the capacity to recognize and respond to moral considerations is central to accountability. Attributability is out of scope for him.

## 2.3 Reasons and moral motivation

Moral reasons for action as an input variable to agential reasons-responsiveness are also an important factor for judging the agent morally responsible. One can be an internalist or externalist about reasons[59]. Internalists think that one can only

[58] Vargas, Building Better Beings, p. 145
[59] https://plato.stanford.edu/entries/reasons-internal-external/, last access: 14.7.2021

respond to reasons if they have them properly internalized. Maybe the dictator's son JoJo can respond to moral considerations in general, but he never noticed that causing pain to others could be morally wrong. According to externalists about reasons, the reasons exist in society and it is not possible not to encounter them. An agent is expected to follow moral reasons even though they never internalized them. This neighbouring philosophical problem is deep but also important for our judgments about moral responsibility of manipulated agents, because some kinds of manipulation are based exactly on disrupting the process of internalization. It is the part of the debate where there are the biggest terminological inconsistencies. Most of the philosophers abandoned the term *reasons*, because it is already very philosophically burdened and established their own terminology for discussing moral motivation.

Vargas prefers the term moral considerations because reason is conceived of as something like an autonomous faculty that properly operates independently of the effects of moral affect[60], at least in the rationalist conception of agency[61]. This seems to imply he wants to focus on internal motivation only.

Vargas explains that internalised moral reasons are a product of sustained exposure to external reasons for compliance[62]. Our reactive attitudes "initially work by providing external motivation for agents to track moral considerations and regulate their behavior in light of them", but under many conditions, we will obey the external motivation and it will eventually become a norm that is experienced as intrinsically motivating[63]. After this short explanation, Vargas stops using the

[60] He is also following the Strawsonian view in understanding moral judgements as something based on emotions. I believe that emotional resentment is a result of learning to recognize evil and it is not very important for my cause, how we experience it.

[61] Vargas, Building Better Beings, p. 203

[62] Harland, Beyond the Moral Influence Theory? p. 408

[63] Vargas, Building Better Beings, p. 175

language of intrinsic motivation, but based on that Harland suggests interpreting Vargas' following concept of *responsiveness to moral considerations* as:

"A person is responsive to some moral consideration *M* to the extent that he is intrinsically motivated to act in accordance with *M*."[64].

Mele does not stay with the notion of *reason* either. He calls moral reasons regularly *desires*, which seems to be an incorrect use of psychological terminology. He explains in his book *Motivation and Agency*:

"Philosophical work on reasons for action tends to be guided by concerns with two distinct but related topics: the *explanation* of intentional actions and the *evaluation* of intentional actions or their agents. In work dominated by the *explanatory* concern, reasons for action tend to be understood as states of mind. Philosophers concerned primarily with *evaluation* may be sympathetic or unsympathetic to this construal, depending on their views about standards for evaluating actions or agents. [...] A theorist who holds that the pertinent notion of rationality is subjective [...] may be happy to understand reasons for action as states of mind [...]. A theorist with a more objective conception of rational action [...] may find it very natural to insist that many or all reasons for action are facts about the agent-external world.[65]"

It could mean that he acknowledges that there is a disagreement about the externalist or internalist nature of reasons in moral theories regarding their evaluative concern. His goal, in contrast, is to *explain* what motivates human behavior and he does not want to use the term *reasons* for it, because of its possible externalistic interpretation. He wants to discuss the internalized motivation only and to not confuse his readers. This interpretation is also supported by Clarke, who mentions that for Mele, the label *desire*, even for something an agent does not value,

[64] Harland, Beyond the Moral Influence Theory?, p. 408

[65] Alfred Mele, *Motivation and Agency*

gives us a thin and subjective notion of rationality and reasonableness[66]. Mele is (according to Clarke) willing to allow that an agent may have reasons that are external, however, he suggests that they can contribute to explanations of actions only by way of the agent's recognizing them and acquiring the sorts of attitudes that causalist action theorists call reasons.

Mele also introduces an umbrella term for desires and non-desires called *motivation-encompassing attitudes.* They are supposed to be desires to *A*, intentions to *A*, beliefs that we ought to *A*, and beliefs that we will *A*[67]. In his later books, including *Manipulated Agents*, he calls them simply *pro-attitudes*.

In the next chapter, I would like to analyze in more detail the control conditions and the historical condition. I will describe different modes in which manipulated agents may fail to have control and I will introduce Perebooms's famous Four Case Argument, which captures most of the failure modes. I consider manipulation a special case of a failure and I believe that if we could see it in a context of different failures, it may reveal some inconsistencies that philosophers commit while unsystematically taking into account some factors and not others.

# 3. The control condition

*Control* could mean several things depending on the model of decision making we have. We can find ourselves in a designed situation, we can get manipulated information, or we can be a specific target of global manipulation, such as those in manipulation cases.

[66] Randolph Clarke, "Motivation and Agency by Alfred R. Mele", *Mind* 113 (2004), p. 565-569
[67] Mele, *Motivation and Agency,* p. 19

Obviously, if we get false information, our control structures are not replaced, but manipulators are giving us the manipulated information at moments when the probabilities that we will check its correctness are low. It means that although we are a fully responsible adult, there are only a few chances we can get out of the situation. Still, we can say that it is our responsibility to know that these kinds of fraudsters exist and that they use a specific, clever trick or a lie to make us do something unethical.

I have already described that control means in *Real Self* accounts something like harmony or inner alignment. The focus is on the degree to which our choice of action is well integrated with the rest of our motivations, so we do not have any opposing tendencies. The next step is to show what Vargas' account has to offer.

## 3.1 The model of responsible agency & Vargas

For Vargas, the question of the epistemic condition being (or not being) a part of the control condition described in the chapter 2.1.1 should be answered by having two distinct epistemic requirements. One is the foresight condition which is more general and is considered a part of the self-directed agency condition (although Vargas mentions that foreseeability of the effects of an action could be also drawn as an independent third item in the figure besides self-directed agency and free will). The other epistemic requirement is recognitional capacity. It is a more particular, free-will specific requirement of recognizing epistemic moral considerations. If we are Mr. Balda, we can have a general epistemic knowledge about religious wars and consequences of being a neighbour to someone with different religious customs. We can be right that increased attention to the topic can help people to take a more protective stance to their customs, but we can miss the particular moral consideration that achieving this by cutting down the tree on the tracks is a crime and it threatens the lives of our fellow citizens.

To treat foreseeability as separate from the free will condition seems to be useful also from the perspective of failure modes. If we fail to consider some particular consequence that means nothing morally, it could be exculpated, but failing to recognize moral considerations for the same action is the main reason for blame. Of course, in reality it is connected: Only if we can foresee events in general, can we, probably, recognize their moral significance. But if we did not see them coming, we can still claim that it was only a case of ignorance and we would have definitely considered the moral side of it if we ever knew. For example, if we lacked epistemic knowledge that someone's foot was nearby and had no reason to think it might be, it is not a matter of ill will, or a failure of due concern by us in light of morality, when we accidentally step on someone's foot[68]. We also may know that it is probable that someone's foot is there (have foresight) and fail to recognize it as a moral reason not to step there (recognitional capacity).

Vargas thinks that we can also satisfy the moral epistemic condition without having reasonable foresight. He claims that "in a particular situation, one might have a good grasp of the salient moral considerations that ought to play a role in one's deliberation, but all the same, one might have little grasp of the likely consequences of choosing one way rather than another.[69]" I do not think it is sufficient, because this kind of explanation offers only potential recognition. Someone can recognize relevant moral aspects because we saw them in other relevant situations, where they also foresaw the consequences, acting morally. Another situation could be that we know that they can recognize moral considerations theoretically, because they talk about it. However, we need a recognitional capacity that actually recognizes relevant situations where moral considerations apply (or could apply), to really act on it. The research of Keith Stanovich shows that it could be way more complicated with the epistemic part of the control condition. His book *What Intelligence Tests Miss: The Psychology of Rational Thought* overflows with examples of people who have excellent

[68] Vargas, Building Better Beings, p. 234

[69] Vargas, Building Better Beings, p.202

thinking skills (potentially) but they perform very poorly in recognizing the situations where they should apply them. He calls them cognitive misers and they could be a significant part of the population. It could be problematic for Vargas and I will go back to it in the context of failure modes later in this chapter.

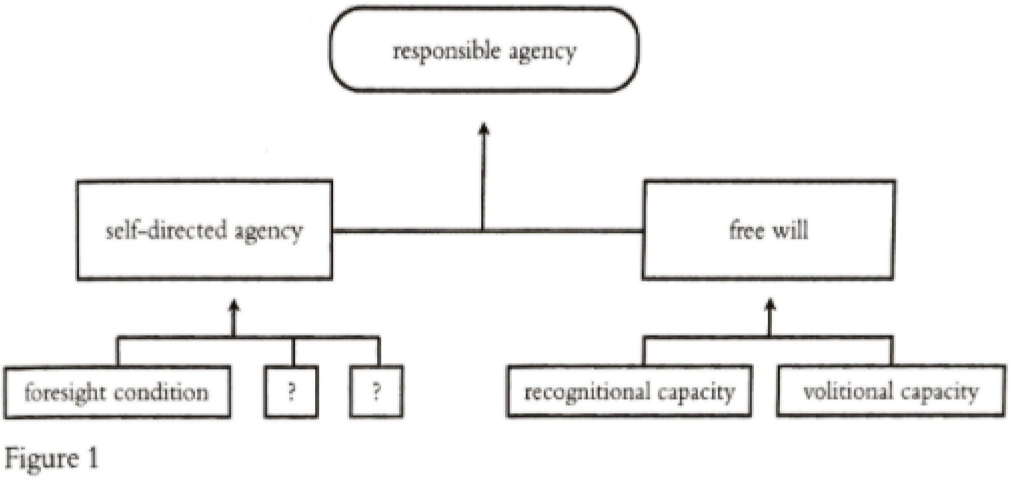

*Figure 2 - Vargas' account*

The figure above shows Vargas' model of responsible agency. He thinks of his account as a prescriptive one which requires some construction, i.e. idealization that is normatively structured[70], but does not postulate any entities that are at odds with a broadly naturalistic view[71]. Harland thinks of it as a two-part test for responsible agency[72], because the whole model consisting of conditions (A and (B) is a capacity sufficient (but not necessary, because we can also have indirect responsibility) for responsible agency. Vargas provides the following description:

"An agent S is a responsible agent with respect to considerations of type M in circumstances C if S possesses a suite of basic agential capacities implicated in

[70] Normatively structured idealization in a sense that he does not want to describe the current moral practice, he wants to set the norms that should be followed in order to cultivate moral agency. But not that idealistic, so it departs from the naturalistic picture of the world.
[71] Vargas, Building Better Beings, p. 216
[72] Harland, Beyond the Moral Influence Theory?, p. 417

effective self-directed agency (including, for example, beliefs, desires, intentions, instrumental reasoning, and generally reliable beliefs about the world and the consequences of action) and is also possessed of the relevant capacity for (A) detection of suitable moral considerations M in C and (B) self-governance with respect to M in C.[73]"

Condition A (the control condition, free will) is constituted by two distinct capacities: recognitional and volitional. Recognitional capacity specifies a detection condition for the control (or free will) and for Vargas it is as follows:

"(A) the capacity for detection of the relevant moral considerations obtains when:

(i) *S* actually detects moral consideration of type *M* in *C* that are pertinent to actions

available to *S* or

(ii) in those possible worlds where *S* is in a context **relevantly similar** to *C*, and

moral considerations of type *M* are present in those contexts, in a **suitable proportion of those worlds** *S* successfully detects those considerations."

The position is highly variable. It does not suppose that there is something like a physically stable realizer and allows that what counts as responsible agency can vary considerably across contexts even for the same agent with no intrinsic change in that agent. Vargas also points out that the detection does not need to be conscious and that the mechanism of this awareness is likely to be diverse and even variably constituted at neurological levels.

## Volitional capacity

[73] Vargas, Building Better Beings, p. 214-215

The second condition (B) called self-governance or volitional control condition on responsible agency is supposed to be a more basic requirement which is not tied to responsibility specifically, but its distinctive part needed for responsible agency could be characterized as follows:

“(B) the capacity for volitional control, or self- governance with respect to the relevant moral considerations (M) in circumstances (C) obtains when either

> (i) S is, in light of awareness of M in C, motivated to accordingly pursue courses of action for which M counts in favor, and to avoid courses of action disfavored by M or
>
> (ii) when S is not so motivated, **in a suitable proportion of those worlds** where S is in a context **relevantly similar** to C
>
> > (a) S detects moral considerations of type M, and
> >
> > (b) in virtue of detecting M considerations, S acquires the motivation
>
> to act
>
> > accordingly, and
> >
> > (c) S successfully acts accordingly.”

To explain the *relevant similarity* in his definition of detection capacity, he gives an example: If we suppose the *relevant similarity* is very coarse-grained, i.e. the bare fact that the context is actional or deliberative makes it the same context, it makes the relevance too expansive and it would not allow us to distinguish between contexts where an agent is trying hard to ascertain the moral considerations and “contexts where, say, an agent was non-culpably extraordinarily time-pressured or subject to manipulation[74]”. It would almost always render agents capable of detecting moral considerations and become very demanding.

On the other hand, if the relevance is too fine-grained in a deterministic world, *similar* would mean exactly identical and then we will have a very permissive

[74] Vargas, Building Better Beings, p. 218

conception (where we cannot effectively blame anyone). This leaves us with a moderate degree of granularity of a non-subjective conception of relevance which Vargas describes as being co-optimal or better for fostering agency that recognizes and suitably responds to moral considerations.

Vargas explains that limit cases such as *Frankfurt cases* and alike are likely to create their own idiosyncratic classes of relevant similarity, i.e., *Frankfurt cases* contexts might have their own rules about what counts as a capacity, quite apart from more ordinary contexts[75].

The notion of a *suitable proportion of worlds* in his formula is explained as co-optimal for a practicable system of judgments, practices, and attitudes shaped by the aims of the responsibility system. However, Vargas agrees that our agency is not ideal and we also never have certainty that all the conditions are met.

He also opposes the *Real Self* views by mentioning: "What does matter is that it is a moral consideration that moves the agent to act either actually, or in the suitable range of counterfactuals. How and whether we judge the springs of our own action is less important than what in fact moves us to act.[76]" He adds that the agent is responsible even if we discover that they misrepresented their motivations to themselves.

Two parts of this model are possibly problematic for me. First, the word *relevance* is a keystone of the definition. All the pressure is on the fact if we count a world where an agent responded to *M* as relevantly similar. The coarse-grained/fine-grained specification does not give us enough clues for where to lead the line. Second, Vargas explicitly refers to the *ideal observer* with his *non-subjective* conception of relevance. He writes that "the notions of suitability and relevant similarity invoked in (A.ii) and (B.ii) are given by the standards an ideal, fully informed, rational, observer in the actual world would select as at least co-optimal for the cultivation of

[75] Vargas, Building Better Beings, p. 220

[76] 217

our moral considerations-responsive agency[77]". These two references are not very illuminating and they leave us with a vague model of responsible agency.

## 3.2 Circumstantialism

In some respects, the vagueness of the agential model could be also understood as its strength. Whether one is self-governed in one or another context partly depends on whether or not the detection condition has been satisfied. This is an important connection making self-governance tied to particulars, i.e. specific situations with specific moral considerations. Vargas claims to be skeptical of there being a single, unified thing that constitutes control across all circumstances, both internal and external to the context of responsibility[78]. For him, responsible agency is a function of the agent and their circumstances, and therefore is highly context sensitive. The general capacity of reasons-responsiveness is meant to be really a cluster of more specific, ecologically limited capacities indexed to particular circumstances[79] and the emerging picture of agency in the social, cognitive, and neurosciences supports this view.

Vargas' circumstantialist picture is one in which an agent's control can vary across context and relatively to the involved moral concern with no variation in intrinsic features of the agent[80]. It contrasts with most of the other accounts of responsible agency, including Frankfurt's, which present one fixed intrinsic feature as the general capacity for general responsible agency and therefore cannot explain variances in capacities of the agents.

In those accounts it is not possible to fail to detect a given class of moral consideration and be nevertheless able to self-govern in light of moral

[77] Vargas, Building Better Beings, p. 214
[78] Vargas, Building Better Beings, p. 223
[79] Vargas, Building Better Beings, p. 205
[80] Vargas, Building Better Beings, p. 226

considerations of another type. If we are bad at detecting considerations of kindness, it does not influence our considerations of loyalty and we can self-govern in circumstances where only loyalty considerations are needed. Or another Vargas' example: “My ability to resist the impulse to make rude remarks to students and strangers need not mean that I am resistant to such impulses when with family members over the holidays.[81]”

Vargas describes other accounts as having the atomistic tendency: responsibility's requirements are understood in a way that focuses exclusively on the agent. He criticizes that it takes responsible agency as a set of properties of an agent in a vacuum and argues that self-governance or self-control should be conceived as notions from social psychology.
His responsible agency model is claimed not to be a **cross-situationally stable mechanism**, because it seems unlikely to him that we have any such thing.

## 3.3 Failure modes of the model of responsible agency

To summarize it, I would like to go through the options of how we can fail to be responsible according to Vargas' agential model. Vargas  is distinguishing two kinds of exculpation: excusation and exemption. We can be excused based on our ignorance in cases where the harm was unintentional (see the example with stepping on someone's foot above). Some agents are, however, exempted from responsibility even before they perform an action, because their reason-responsive structures are impaired.

**Exemptions**

[81] Vargas, Building Better Beings, p. 225

Vargas provides a few examples of exemptions where he does not specify which part of the responsible agency was impaired, but it seems that both epistemic conditions (foresight and recognition of moral considerations) for reason-responsiveness are impaired. One example is a soldier who suffered a head injury and is not able to recognize reasons. If he is commanded to kill prisoners, he is thanks to his predisposition exempted from being held responsible[82].

The important thing to note is that Vargas considers indirect responsibility being a valid part of social practice. It means that there could be cases in which we are not currently reason-responsive, but at the time we chose the actual course of action we were reason-responsive. We can imagine that the soldier was a healthy, reason-responsive human at the time he received the order to kill the prisoners and he decided to get drunk so "that he becomes numb to moral considerations generally.[83]" Another example are so-called tracing[84] cases featuring drunk driving. According to Vargas, we are morally responsible in these kinds of situations.

### The locus of failure

Those could be examples of an impairment of the reason-responsive capacity in general, but Vargas thinks that agents might be responsible in some circumstances, owing to their possession of the appropriate sensitivity and responsiveness to some minor moral consideration, but not in other circumstances. Example of this could be the mentioned case of responsibility to loyalty without responsibility to kindness.

[82] Vargas, Building Better Beings, p. 272

[83] Vargas, Building Better Beings, p. 272

[84] Tracing means that their moral responsibility traces back to the time when they were sober and did the decision to drink.

The problem with the sensitivity to certain classes of moral considerations is that Vargas claims that we do not count as responsible only if we have a **complete lack of sensitivity**. It implies sensitivity to particular moral considerations as a scale, which seems to be an upgrade in comparison to other accounts where the detection capacity or the whole capacity for moral agency is treated as a binary term. It is a sad thing to construct a very flexible account model of a morally responsible agency and still being unable to assign blame relatively to the advancement of the capacities. However, it could be also interpreted that being morally responsible means at least some moral responsibility. In that case it makes sense that only a complete inability to detect moral considerations exempt agents from moral responsibility and it also explains why Vargas judges most of the manipulated agents responsible.

Besides content- and circumstance-specific failures of recognitional capacity for moral considerations, there could be a failure of a specific part of the agential model. Vargas provides an example of a failure mode where an agent fails to be governed in light of *M*. In case of failure situated in the volitional capacity, the agent recognizes the moral consideration, but it has no proper connection to the production of the action and that means that B(i) is not fulfilled. Consider Diego who has a habit of providing bandages to strangers. He does not do it to help others, it is his ritual executed every Tuesday at noon. Diego recognizes that a particular tourist needs a bandage, but he gives it to them because it is Tuesday, not because they need it. There is a need for a connection of content between the moral consideration and the course of action and it is not present in Diego's case[85]. The problem with the case might be that if Diego is not praiseworthy for his deed, he would not be blameworthy for the same reasons if he was ritually punching tourists in their heads (without ill will and without connection to the right moral

[85] Vargas, Building Better Beings, p. 224

consideration). It does not seem right, but Vargas can object that those are folk intuitions which should be replaced by his normative model.

A case of partial exemption could also be acquired psychopathy. Vargas also claims that acquired psychopathy as a result of manipulation is a valid excuse from moral responsibility[86]. Harland notes that this is inconsistent with his prospective account of blameworthiness and the process of internalization: “Even if psychopaths lack moral sentiments, many are able to adjust their conduct in light of prudential concerns. It seems plausible to think that repeated exposure to negative stimuli (e.g. blame) could give at least some psychopaths an affective drive not to engage in wrongful conduct. According to ACM [Vargas’ agential cultivation model], this should make them moral agents.”

These failure modes may seem sufficient to determine moral responsibility in the manipulation cases, yet Vargas did not see any of his failure modes as suitable to accommodate manipulation. It will be shown on the Four-case argument.

## 3.4 Character control & Mele

As opposed to Vargas, who is using the word “control” in a specific sense of *being able to respond to morally relevant considerations and self-govern in light of them*, Mele writes in his book *Manipulated Agents* about control in a more self-constituting sense. For him, control is a behavioral practice during which we govern ourselves to grow into a person we want to be - as opposed to our childhood, when we are “inevitably fashioned and sustained, after all, by circumstances over which we have no control”[87]. Characters from his examples (Sally and Chuck introduced in the next chapter) are depicted as adults with their own active role in their self-development.

[86] Vargas, Building Better Beings, p. 278
[87] Mele, Manipulated Agents, p. 92

As Mele writes, it is relevant for moral responsibility when agents' capacities for control over their mental lives are being bypassed[88], which is what is happening in manipulation cases.

This is Mele's set of examples of exercising capacities for control over one's mental life and one's consequent character:

> "Sometimes we are told that, or wonder whether, we care too much—or too little—about our work, what others think of us, our children's success, how we dress, money, our health, or whatever. Sometimes, on reflection, we judge that we should care less—or more—about some of these things. Occasionally we make efforts to get ourselves to care less—or more. Someone who becomes convinced that he cares way too little about his health may try to get himself to care much more by spending time each day picturing opportunities that would be closed to him by poor health and thinking about the ways in which better health would improve his life, someone who judges that he cares way too much about work may attempt to fix that by reflecting periodically on the good things that his work leaves him little time for, and so on. Sometimes such efforts are successful, and no such effort would succeed if the values at issue were unsheddable (at the time). [Values that] are revisable in this way . . . differ from unsheddable values that some workaholics, misers, and health fanatics may have. How, exactly, the distinction is spelled out by a particular compatibilist depends on that theorist's preferred way of understanding what it is to have been able, in deterministic worlds, to do things that one did not do. I leave that open[89]."

I mentioned bypassing our control by skewing our detection capacity to be insensitive in particular situations but this is another kind of bypassing. If we want

[88] Mele, Manipulated Agents, p. 45
[89] Mele, Free Will and Luck, p. 186

to manipulate an agent to act against their unshedable values which constitute their character, we need to manipulate them in a more extensive way. Mele shows in cases called *Radical reversals* that if an agent worked hard on their character to become a nice person, and it was reverted forcefully by some kind of (neurosurgical) intervention, it is a good reason to exclude them from moral responsibility.

This means a more serious threat to *Real Self* accounts, namely to Frankfurt's structural condition for being morally responsible. There could be some kind of disharmony in agents who were manipulated to be insensitive to moral considerations, in particular situations in which they are not in other relevant situations. This asymmetrical sensitivity cannot count as wholeheartedness and it will exclude them from responsible agency. But if their whole character was manipulated there could be a case where an agent has no reservations to his act and we cannot tell from his internal structure that he was manipulated to be as he is.

So far, I have introduced three approaches to the control or free will condition, which is the crucial one for moral responsibility. (1) Frankfurt's structural conditions in his *Real Self* account, (2) Vargas' model of responsible agency were already discussed. (3) Mele's historicist approach based on a negative condition will be discussed in chapter 4. I also mentioned some factors that can play a role in judging the agents responsible and some failure modes in which we can do the wrong thing. This could be explained further in Pereboom's famous *Four-case manipulation argument*.

## 3.5 The Four-case argument

Pereboom's *Four-case manipulation argument* is great for detailed analysis, because it is well commented on by various philosophers who revealed their intuitions about it and it shows many possible failure modes to analyze. The four parts of the

argument[90] describe four stories featuring Mr. Plum, an egoistic person in a deterministic universe, who decides to kill Mr. White. The assumption is that Plum is causally determined to do that by different versions of the circumstances of his life. We also suppose that he would not do it without the circumstances described. All the cases mention that his behavior matches his often selfish character and that he does not act out of an irresistible desire, but his decision results from a reason-responsive process of deliberation.
It goes as follows:

1. Plum is manipulated by evil neurosurgeons to kill White. He is an egoistic person, but usually succeeds in regulating himself. In this situation, the neurosurgeons were manipulating him by radio-like technology, specifically by pressing the button just before he started to reason about his situation.
2. Evil neurosurgeons manipulated Plum at the beginning of his life to make his reasoning often egoistic. In his situation, given his programmed egoistic reasons, he decides to kill White.
3. Plum grew up in a community which trained him so that his deliberative reasoning is often egoistic. His acquired tendency to engage in egoistic reason-responsive processes leads him to kill White.
4. Plum was raised in a normal family and because of his developed egoistic tendency in the deliberation process, he kills White[91].

The pressure is to draw the line between responsibility and non-responsibility and make the distinction a principled one. According to Pereboom, Plum is not different in the four cases in any way relevant for responsibility. All the Plums are like a normal agent in normal circumstances: causally determined. The circumstances of

[90] Derk Pereboom, *Free Will, Agency and Meaning in Life*. New York: Oxford University Press, 2014. ISBN 978-0-19-877686-4. p. 77

[91] ibid., p. 76-79

how Plum was determined to kill do not matter. If we think that Case 1 is not a case of responsibility, we should think that about all the cases[92].
In the description of the cases, Pereboom makes sure he describes Plum meeting most of the possible conditions for  free responsible agency described by *Real Self* accounts (and other accounts which bothered to define an internal condition for moral responsibility): Plum's first order desire conforms to his second order desire. That means he does not have a desire not to have the desire to kill White. Also, no irresistible desire on an emotional level which would override reason-responsiveness plays a role. Plum decided based on his rational egoistic reasoning as he normally does, just this time, he was manipulated to pay more attention to the egoistic reasons than he would in the same situation and it tipped the scales resulting in another final decision. Plum's action does not conflict with his character, he acts often egotistically in other situations, though not always.

As Pereboom explains in his replies to objections, the argument features manipulation just to make it intuitive that something makes an agent unfree. "The next step is to argue that non-responsibility is preserved even when the manipulation is subtracted, on the ground there is no responsibility-relevant difference between the deterministic case that features manipulation and one that does not.[93]"

Although the argument originally serves to show that we have no control at all, it shows us a lot about what philosophers really mean by their accounts.

## 3.6 Reactions to the Four-case argument

[92] Ibid.
[93] Ibid. p. 80

The argument has not been accepted by historicists or structuralists. Mele is refusing it by claiming that causal determinism is not the best explanation for the intuition that Plum is not responsible in Cases 1-3 as Pereboom says. "So far, at least, the claim that the manipulation in those cases is what does the intuition-producing work looks very plausible,[94]" he writes. He also provides alternative arguments with indeterministic scenarios which have the same effect with no determinism involved to support his thesis, but we have already ruled out indeterminism from our playing field.

In structuralist positions, some philosophers claimed that they have no intuition about Plum not being responsible in Cases 1-3[95]. Compatibilist McKenna gave the hard-line reply saying that since the normal deterministic agent is morally responsible, so is Plum[96].

Stephen Kearns objects much like Mele does that manipulation is what makes the first cases intuitively about non-responsibility and it does not transfer to the ordinary case[97].

An interesting answer comes from Fisher[98]. He claims that there is a difference between responsibility and blameworthiness as I already mentioned in the previous chapter. Consistently with his view, he claims that there is no difference with respect to the minimal control conditions for moral responsibility, making Plum responsible in all the cases. The minimal control condition for him is the *guidance control* I mentioned in the first chapter and is defined in terms of "action flowing from the agent's own, moderately reason-responsive mechanism[99]". However, there are wide disparities in the conditions for blameworthiness including the

[94] Mele, Free Will and Luck, p.141

[95]Harry Frankfurt, "Reply to John Martin Fischer", in: Sarah Buss and Lee Overton (eds.) *Contours of Agency: Essays on Themes from Harry Frankfurt*, eds.. The MIT Press. 2002. 27–8

[96]Michael McKenna, "A Hard-Line Reply To Pereboom's Four-Case Argument", *Philosophy and Phenomenological Research 77* (2008), p. 142-159

[97] Pereboom, Free Will, Agency and Meaning in Life

[98] John M. Fischer, "Responsibility and Manipulation", *The Journal of Ethics 8*, (2004), p. 158

[99] John M. Fischer and Mark Ravizza, Responsibility and Control: A Theory of Moral Responsibility (Cambridge University Press, 1998).page 207

circumstances of the creation of Plum's values, character or desires. He also says that there is no reason to suppose that anything like such unusual circumstances exist only as a result of causal determinism.

Vargas considers Pereboom's Four-case argument a problem for our commonsense conceptions of freedom and responsibility and claims that his normative account is different.
According to him, Plum has the capacity to recognize moral considerations and is therefore responsible even in Case 2. In Case 1, Vargas claims that details matter. If the neuroscientists push buttons to manipulate Plum to start rationally egoistic reasoning, they may do it *by altering inputs* while leaving his deliberations and mechanisms of rejection intact. Only if the agent's reasoning capacity and deliberation mechanisms are micro-managed and flawed, will Vargas admit that the agent is not responsible[100].

**Inputs and structure**

As he explains, the replaced control structures work more as an excuse for agents than if they were able to control their actions and were subjects of manipulated inputs only. But what does this suddenly appearing distinction mean? The only specification Vargas provides is with desires. If we have some additional desires (inputs) and their strength doesn't disrupt our ability to consider others in our behaviour, then we still have the kind of control that suffices for moral responsibility. He adds, cynically, that if we think that we lose the ability to respond to moral consideration and it disrupts our responsibility after receiving a new desire, we should think the same thing about any food advertisement that makes us hungry[101]. But the question is when it is disrupted and if disruptive desire counts as replaced control. Vargas says that there could be a really strong but not irresistible

[100] Vargas, Building Better Beings, p. 287
[101] Vargas, Building Better Beings, p. 277

desire, which may disrupt control, but not responsibility-relevant control. Or there could be cases where our responsibility-relevant control will be maintained, but other control disrupted. But it does not bring us any closer to some specific examples of replaced control structures. I can guess that ordinarily replaced or impaired control, where we fail to apply the right considerations, could be the case, for example, when intoxicated. Or maybe we should interpret it that disruption of responsibility-relevant control by a strong desire means the replacement. In any case, it still does not help us in cases of manipulation by evil neurosurgeons, where the control structures are in place, but our values are manipulated.

I would like to oppose the distinction between structure and input manipulation. At least in some models of decision-making, there is not a big difference between inputs and the control structures themselves. The repetitive inputs are just promoted to "the position of structure". All the models of learning based on reinforcement rely on inputs, reinforcement of the inputs by repetition and reward. There are also association theorists who believe that there is no structure in our knowledge besides the one created by inputs which were commonly taken together[102]. This means that it is even harder for us to distinguish between a harmless advertisement and manipulation. The same ad could be harmless for one agent and manipulative for another, less resilient one, who has already seen a lot of similar ads and is about to create a destructive opinion based on it. It also means a good thing: we are not even close to overnight manipulations, because radical reversals need time for proper reinforcement.

### The role of informations

[102] What is Associationism? in Associationist Theories of Thought https://plato.stanford.edu/entries/associationist-thought/#WhaAss accessed 5.8.2021, First published Tue Mar 17, 2015; substantive revision Wed Jun 24, 2020

The counterfactual information that Plum would not decide to kill White without the intervention (whatever it is across the cases) seems interesting to me. This is something unique to theoretical scenarios. We never have this type of information while judging real-life scenarios from our bounded point of view and so far, only Mele and his students are considering seriously indeterministic scenarios where the manipulation consists only of increasing chances for some specific outcome. But what we could have is information about the externalities. Given Plum's character, nobody but his mom would suspect that anything extraordinary happened to make him commit murder. However, information about the manipulation changes the game. It shows an interesting relationship between our knowledge about mechanisms in the world (influence, internalization, harmful social practice) and our actual judgement about moral responsibility of the manipulated agent: Based on the fact that we will never know how far Plum was from killing someone, we can only deduce from what we know about other people growing up in an environment practicing deviant social engineering to what degree their environment is responsible for their current condition. We will actually imagine what an ideal observer could have known based on our actual knowledge and exculpate Plum to the degree we think the environment participated in his decision.

## 3.7 The roots of the disagreement

**Prospective versus retrospective views**

By the *Four-case argument*, Pereboom wants to show that moral responsibility in the *basic desert sense* is in conflict with the basic deterministic rules of the physical world. That is because the *control* in action required for an agent to be *truly* deserving of blame and praise is a kind of control we never have. He holds that there are other senses of moral responsibility that are plausible. In chapter 6 of his book, he suggests a concept of blame as an interhuman practice to prevent immoral behavior and its consequences, i.e. a process where addressing wrongdoing is a stage to moral formation of an agent or their (or others') realization of a particular objective of moral consideration[103]. This approach is actually similar to Vargas' although Vargas claims that his conception is different from "pereboomian". There are at least two important factors which they have in common. Both are considering moral responsibility as social practice and both are trying to redefine deserts in forward-looking terms as I already explained in chapter 2.1.

As I mentioned, prospective interpretation of blame is a bit controversial. John Doris[104]criticizes it as follows: the common understanding of desert, even supported by Oxford English Dictionary, is backward-looking. Its retrospective function is essential and if Vargas defines desert as a prospective term with a function to make better people, he does not do the revision, he is stipulating a brand new term. For Doris, it does not make sense even after Vargas' upgrade of the concept (now Vargas claims that it has two levels: particular judgements are backward-looking, but the whole responsibility practice is forward-looking). Doris illustrates it in the example of sports. According to Vargas a player should say: "We deserved to win the title, because it will make me a better player." But according to the theorist defending backward-looking desert the sportsman should say: "We deserved to win the title, because we fairly won 4 of 7 games in the final series."

It could be easily objected that this is not the practice Vargas has in mind and the forward-looking practice can include backward-looking blame: The sportsman may say that "We deserve the title, because everyone saw that we are the best and it will motivate everyone to also do their best." But one can still notice that the backward-looking desert is somehow closer to the notion of fairness than the forward-looking one. Doris says that the consequence of forward-looking blame could result in that the most incorrigible offenders would deserve the least

[103] Pereboom, Free Will, Agency and Meaning in Life

[104] John M. Doris, "Doing without (arguing about) desert", *Philosophical Studies volume 172 (2015), p. 2625-2634*

punishment and it also makes it possible to punish someone if it is likely that they would benefit from it. Does it help if retrospective and prospective practices are situated at different levels of moral discourse and the particular judgements are backward-looking while the whole practice is forward-looking? While we can tell that punishing particular agents is retrospective and the whole practice is prospectively justified, the benefit of the whole practice always breaks down to the particular judgements with the consequences I just mentioned. Doris wraps up with a point that we still have an unresolved substantive and normative dispute regarding the two level psychology of desert. I will wrap up that we have another factor influencing whether we will judge a manipulated agent morally responsible or not: if we have a forward- or backward-looking conception of desert. Or more precisely: If we consider the benefits as primary in our moral practice and fairness as secondary or if it is the other way around. Vargas seems to appeal more to the beneficial part and therefore he judges the manipulated agents often responsible as opposed to Mele whose intuition would be more aligned with fairness and he prefers not to blame manipulated agents.

To sum it up, the argument divided philosophers (or at least compatibilists) into two camps: Those who are taking determinism seriously gave a hard-line reply (that Plum is responsible in all the cases) and are interpreting the Four-case argument in a way that responsibility actually transfers from the ordinary case to the others and not the other way around. For the second camp, the argument is orthogonal, because their definition of control has nothing to do with the kind of control required in the basic desert sense. In the light of the reactions, it seems that basic deserts are a kind of strawman nobody wants to defend. Positions of philosophers developed and to imagine control as a kind of magical indeterministic sourcehood of our action in us seems obsolete. In the second camp would be two kinds of philosophers: those who redefined moral responsibility as beneficial social practice such as Vargas and those who see ascriptions of moral responsibility as a convention not necessarily dependent on our assumed internal structures (as interpreted by Mele).

## 3.8 Responsibility of manipulated agents

It seems that in manipulation cases, the control is bypassed. But which part of the conceptual construction is not fulfilled? It seems not clear, although a good model of responsible agency should explain it.

In Case 1, Vargas claims that details matter. If the neuroscientists push buttons to manipulate Plum to start rationally egoistic reasoning, they may do it *by altering inputs* while leaving his deliberations and mechanisms of rejection intact. Only if the agent's reasoning capacity and deliberation mechanisms are micro-managed and flawed, will Vargas admit that the agent is not responsible[105].

But this is problematic, as I argued earlier in this chapter. There could be no difference between inputs and the deliberation capacity itself. Furthemore, it does not make sense for Vargas to make this distinction when he already made a claim that we can be morally responsible in a certain context while not being responsible in another. That basically means that we have “multiple responsibility capacities” distinguished by their “aboutness”. Pereboom claims that the neurosurgeons enhanced Plum’s egoism. If we translate this into Vargas’ terminology, it could mean that by increasing his egoism they manipulated his detection capacity for altruistic considerations of kindness or respect. His other detection capacities (for loyalty for example) could have stayed untouched, but those are not the relevant ones for the case. In this interpretation, where inputs are the particular moral factors to consider (e.g. I should behave nicely to Mrs. White), it does not make sense to separate them from their capacity.

[105] Vargas, Building Better Beings, p. 287

In another interpretation, we can imagine the inputs being the reasons for the act of murder.

Pereboom writes that Plum did it for the sake of some personal advantage. Manipulators could have made him see the advantage big enough to proceed with the murder. This is closer to real manipulation scenarios. Plum could have been targeted because his egoism is already high and it would not take too much work to trigger him to act. The same way in which political advertisements target neighbourhoods that are already in serious trouble, even Mr Balda from the example in the introduction could have been targeted by online advertisements based on his age, political preferences and the number of foreclosures imposed on households in his area. Does it make a difference if we do not change a person's sensitivity to moral considerations but instead strengthen the input (or weaken the input that White is actually worth moral considerations)? Now we are again close to paternalism. Someone could say that they just took advantage of the fact that Plum is egoistic and created a situation strong enough for him, so he reveals how low the level of his inhibitions to actually kill someone is. However, Vargas himself appeals to relevant circumstances. Why does he not count that the situation was artificially introduced in any case and it would not happen without the neurosurgeons? Maybe two weeks later, the method the neurosurgeons are using would be revealed in the media and he would take steps to never be a victim of them. I am inclined to say that more details about the neurological process, which Vargas demands, would not help us. What we need is a more detailed model of responsible agency, which includes how our capacities are made or changed. Or at least a better picture of relevance, because every agent has their limits and it does not make any sense to hold agents responsible in the situations relevantly similar to the extent of classes of moral considerations playing a role, but take only a complete lack of detectional capacities as exemptions.

Also a better concept of internalisation is needed which could explain if it is even possible for a capacity to be manipulated in a different way than through its inputs. If there is not such a difference, Vargas will probably hold that manipulated agents are always responsible, because if agents could be manipulated, they could also learn from blame.

However, it also may be beneficial to consider brainwashing as numbing our capacity to recognize reasons and support its victims against the aggressors. One of the typical propaganda tools is to change our perceptions of certain groups of people or to strengthen a particular type of our considerations in favor of another. For example, loyalty considerations are reinforced at the expense of considerations of kindness. An intense brainwashing flaws our detection of moral considerations in circumstances related to a particular groups of people and it can even lead to murder if the manipulator manages to properly dehumanize the groups in someone's eyes. The manipulated agent would then claim that context with the dehumanized group is not relevantly similar to situations where kindness applies and there are already cases of lawsuits where people claimed to be brainwashed[106]. If we start looking at manipulation as a technique for numbing certain types of moral considerations and count the percentage of its influence instead of exempting only the complete lack of sensitivity, we can end up with more precise and more beneficial moral practice.

We can also adopt another strategy which could save us from these kinds of problems. That is why Mele's historicist account (explained in the next chapter) with the emphasis on external factors seems appealing. It makes the fuzzy cases of manipulation more workable if we redirect our attention from complicated undetectable conceptual structures to external situational factors and say that the

[106] Lawyer for accused Capitol rioter says client had 'Foxitis,' 'Foxmania'. https://thehill.com/homenews/media/552285-lawyer-for-accused-capitol-rioter-says-client-had-foxitis-foxmania last access: 6.8.2021

presence of manipulation itself is a factor making the manipulated agents not-responsible.

# 4. Historical condition for moral responsibility

The discussion about the conditions for moral responsibility was focussed mainly on the formulation of the internal condition. However, there are philosophers attempting to include or otherwise consider an external condition in their account, namely Mele and the team of Fischer and Ravizza. The accounts which are trying to include it are called historical, because they suppose that an agent needs to have the right kind of history to be morally responsible - as opposed to structuralist accounts arguing that we only need to explore agential psychological structures to decide responsibility. It seems the differences between the accounts create a fruitful discussion in which some unspoken assumptions about the philosophers' concepts of moral responsibility could be revealed.

Arguments for the need of external conditions are of two kinds. One kind appeals to character control and freedom in the moral development of an individual. While it is hard to define some standards for upbringing and education, the assumption that there is a need for an individual to have some room to choose their own personality seems to be plausible. Manipulation arguments of this kind often employ a covert intervention of nefarious neurosurgeons or evil psychologists which is supposed to change the agent's value system against their will. The unusual radicality of the intervention is to bolster the inevitability and non-consensuality of the acquisition of the new psychological structures. Because the artificial structure in manipulated agents is the same as in normal ones, it seems to undermine the structuralist position that the according structure is a sufficient condition for moral responsibility. Manipulated agents will decide to act according to the internal condition, but still, they won’t be morally responsible, because of the manipulation in place.

A slightly different subgroup of these cases, which may also serve to point out that something other than the agential psychological structure at the moment matters in assigning moral responsibility, are divine design cases. The difference is that some sophisticated being can predict what will happen in the future and prepare initial conditions to make sure that an agent will do the desired thing without any invasive action. This is the type of a God-like manipulation as mentioned by Franklin and I was introducing it with the Zygote argument in the first chapter.

The other kind of arguments showing the need for the external condition for moral responsibility features so called tracing cases. In cases where we can trace the sequence of events back to the decision for which the agent is morally responsible, it doesn't matter what his psychological structure is at the moment of the action. Those cases typically feature an agent deciding intentionally to shut down their control while knowing the risks of doing so. I mentioned them in relation to indirect responsibility in Vargas' account.

## 4.1 Classifications

### Externalism about conditions for responsibility & Mele

In his book Responsible Agents, Mele offers a classification, in which we can have an internal condition and on top of that an external one, a historical one. Mele explains conditional internalism in Frankfurt's position:

"An agent's internal condition at a time may be understood as something specified by the collection of all psychological truths about the agent at the time that are silent on how he came to be as he is at that time."

It is a sufficient condition. Then he explains conditional externalism as a position saying that conditional internalism is valid but sometimes agents are responsible because of how they came to be in the internal condition, i.e. an event prior to the action we are responsible for, e.g. we decided to drink before driving. Another causal route can lead to internal conditions, normally associated with responsibility, where an event on the route causes that the agent is not responsible (see Fig.3.).

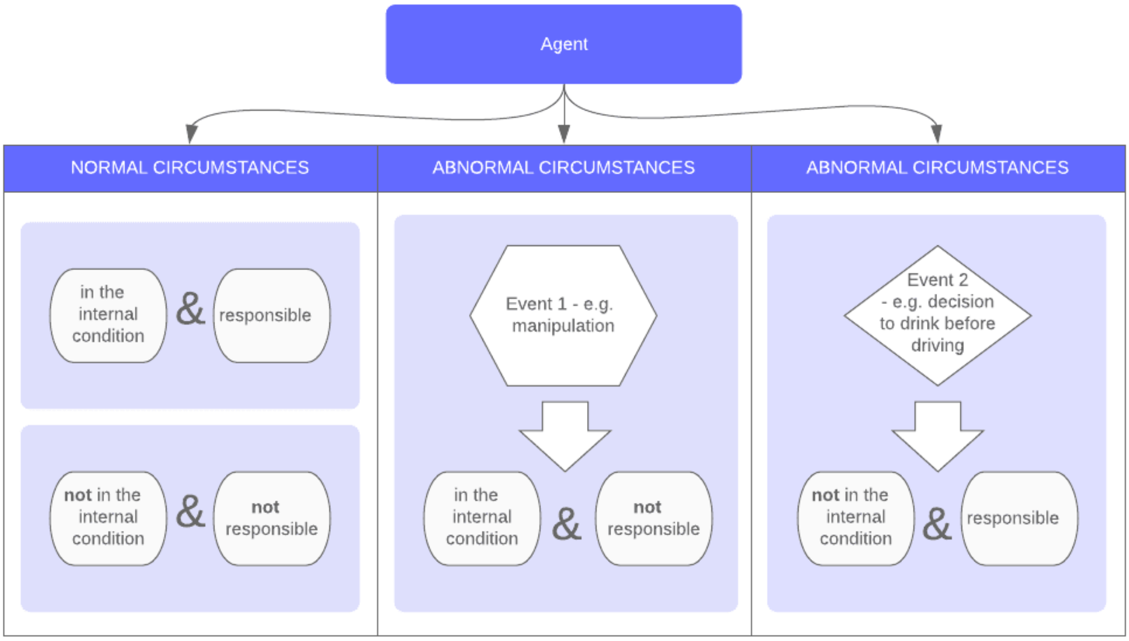
Agent
NORMAL CIRCUMSTANCES
ABNORMAL CIRCUMSTANCES
ABNORMAL CIRCUMSTANCES
in the internal condition
&
responsible
not in the internal condition
&
not responsible
Event 1 - e.g. manipulation
in the internal condition
&
not responsible
Event 2 - e.g. decision to drink before driving
not in the internal condition
&
responsible

*Figure 3 - Combinations of the internal condition and responsibility according to Mele*

To understand Mele's defence of conditional externalism, we can show it in contrast to Vargas' position. Vargas is not an internalist, but he denies the middle part of our figure and claims that manipulation is not an exception from moral responsibility.

### Historical condition & Vargas

Vargas divides the approaches in a different way: as essentially historical, essentially structural (or non-historical) and mixed. An essentially structural view is one where the agent's responsibility does not depend on any facts about their history. In contrast, on essentially historical views, there is always some historical condition that must be satisfied for an agent to be responsible. "It matters how the agent came to whatever structural features he or she possesses[107]". All the views in the middle he calls "mixed".

Vargas declares himself to the mixed position. For him, in some cases structural conditions will be sufficient, but in others, there will be some historical requirement. He suggests that we can think about it also as "variantism" about responsibility depending on some unified story and mentions two reasons to choose the mixed view. First, it derives from his general approach where he focused on providing the structural conditions that are ordinarily sufficient for moral responsibility. However, he claims it is compatible with the above that there may be historical conditions that independently suffice for free will and moral responsibility. Second, he wants to make a concession to "tracing" cases (right part of the figure 3). He says: "A common feature of responsibility practices involves tracing responsibility for an action past the immediate structure of agency back to some earlier point in the agent's history. Drunk driving is one example."
Based on this explanation, it is clear why Vargas did not include historical condition among the other three factors (reasons, identity and control) playing a role in our intuitions about manipulation cases (mentioned in the chapter 2.1). He thinks of it as an exception to his otherwise structural view. But it does not make the historical condition less formative for our intuitions about manipulation cases.

## 4.2 Value manipulation

In *One Bad Day*[108], a manipulation case by Mele, kind Sally is turned into a value twin of thoroughly bad Chuck by a team of psychologists. As a result, she kills her

[107] Vargas, Building Better Beings, p.268
[108] Mele, Manipulated agents, p.20-21

neighbour, George. Chuck and Sally are described as people who worked intentionally on their character. While Chuck's "heart hardening" project included torturing animals, bullying, and killing vulnerable people without mercy (e.g. a homeless man Don), Sally worked hard to grow up from a petty teenager into the gentlest person on Earth for whom doing harm to anyone was not even an option. While Sally slept, Chuck's values were implanted into her system. On *One Bad Day*, she woke up, surprised by her desire to kill George. She found him unpleasant before but never considered any violence. Sally reflects on her desire and her new system of values only supports the desire.

After reflection, she satisfies Frankfurt's structural conditions for being morally responsible:

- She "has no reservations about her desire" and "is wholeheartedly behind it[109]".
- Her desire is "well integrated into [her] general psychic condition[110]".
- Her first-order desires match her second-order desires[111].

She decides to kill him and does so. The next night, the evil psychologists undo everything they had done to her.

Mele explains that the killers, Sally and Chuck, differ markedly in how they came to have their relevant desires and attitudes, and it also makes Sally a victim of external forces. According to Frankfurt's structural account, history makes no difference and only the internal condition defined by the bullet points above is relevant for moral responsibility. Mele objects that while we are inevitably fashioned by circumstances over which we have no control, it doesn't mean that we have no control at all. Chuck

[109] Frankfur, Reply to John Martin Fischer, p. 2

[110] Frankfurt, Reply to John Martin Fischer, p. 27

[111] That means she has the desire to kill George and also a second-order desire of wanting to kill George (as opposed to conflicting first- and second-order desires in which case she would like to get rid of her desire to kill George).

is morally responsible for an extra item – becoming an evil person, while Sally excercised no control in the process of acquiring the Chuckian system of values.

This seems to be a counterexample to structural conditions of any kind, because we can presuppose that manipulated Sally can have any psychological structure imposed on her against her will, not only the one defined by Frankfurt. Mele concludes that in order to be a morally responsible agent, there is also a historical condition - to have a history without manipulation.

There is a similar case introduced by Fischer and Ravizza.[112] They claim that the agent must be responsible for the mechanism that led to the action and a manipulated agent decides based on a mechanism he has no ownership over.

## 4.3 One Good Day

One modification of the One Bad Day case introduced by Mele is One Good Day. It features thoroughly bad Chuck after his heart-hardening project who has no values at all that could motivate a charitable deed. And exceptionally sweet Beth who worked hard on her charitable character.

> "Overnight, without Chuck's consent, they erase his bad values and replace them with good ones that match Sally's. Shortly after he awakes, he starts working with a local Habitat for Humanity crew in his neighborhood. When the workday ends, he drives around town for an hour and buys several boxes of Girl Scout cookies from every Girl Scout he sees—about fifty boxes in all.

[112] Review: Précis of Responsibility and Control: A Theory of Moral Responsibility Author(s): John Martin Fischer, Mark Ravizza, John Martin Fischer and Mark Ravizza Source: Philosophy and Phenomenological Research, Vol. 61, No. 2 (Sep., 2000), pp. 441-445 page 442

> Then he delivers the cookies to a local homeless shelter. His motives are pure, as Sally's are when she does her charitable deeds.[113]"

Consistently with his verdict about post-transformation Sally, Mele considers Chuck not responsible for his good deeds. Vargas's view on the other hand has no clause that rules out such value-transformations, so he claims that post-transformation Chuck in *One Good Day* as well as post-transformation Sally in *One Bad Day* are morally responsible. With appeal to the explanatory power of his account, he holds that dramatic manipulations that are identity-and-moral-considerations-responsiveness-preserving should not count as undermining responsibility.

### The hidden assumption of fairness

Mele holds that his intuition that Chuck is not morally responsible for his good deeds is not misleading based on his radical reversal suggestion. For him, the effort spent on one's character matters and Chuck's pre-transformation character was sufficiently bad that charitable deeds were not even an option for him[114]. I believe that this reveals hidden assumptions about the importance of fairness in his model of responsible agency. Mele says that "the facts about his history that account for his moral responsibility for that character"; and "the facts that account for the good deeds at issue[115]" suggest that the history at issue is not just something revealing some unfair and non-consensual invasive character intervention. As I interpret it, it is also about who did something for the change to happen. This way Chuck received his good character for free and without any effort. Mele does not mention it explicitly, but from his formulations, the hidden appreciation of an effort to make

[113] Vargas, Building Better Beings, p. 29

[114] Mele, "Moral Responsibility and History Revisited," p. 473.

[115] Mele, Manipulated agents, p.51

something happen is what should have an influence on the resulting moral responsibility for an action.

Vargas objects that the effort spent on his previous character is not the explanation of the intuition that post-transformation Chuck is not responsible for his recent good deeds, but of the intuition that Chuck was responsible for the bad things when he did them. That means it is not telling one way or another about post-transformation Chuck. Vargas demands a reason for thinking that those considerations about pre-transformation properly explain anything about the post-transformation Chuck, because it may be just habit or prejudice fueling them. We know that Chuck is bad, we are used to blame him and we can hardly imagine that he is a good person out of the sudden. There may be a reticence to view Chuck as praiseworthy, because regarding his wrongs in the past, he does not feel guilty and he did not try to right his wrongs. Vargas is right that the situation seems conflicted. According to him if Chuck really is recognizing moral considerations and self-governing in light of them, he really is praiseworthy for just that reason. If he is recognizing them and is consistent, he will also regret his past crimes, which is not mentioned in the example. Vargas thinks that we're pretty bad about knowing when and how to praise, implying that even a bad person should be praised for a good deed and I can only add that there could be some contra-productive social pressure against praising someone known for his bad deeds. On the other hand, the interpretation can also be that we know when to praise and how, but we have different goals in praising.

Vargas and Mele agree on the facts that account for the good deeds at issue, but they disagree on whether they tell us anything about Chuck. Mele thinks that the facts, i.e. the role of the manipulators, undermine responsibility. Vargas thinks the only considerations Mele provides for the accuracy of his intuition that Chuck is not responsible turn out to be the intuition itself. For me, the underlying concern about fairness is clear. Manuel Vargas has an idea to interpret it that one can only be morally responsible if one is responsible for one's character, but I consider the aspect of a fair appreciation or condemnation of a moral effort to be more fitting. It is in accordance with backward-looking conception of blame to consider fairness in our ascription of moral responsibility and it has a strong explanatory power regarding our non-responsibility intuitions in manipulation cases. It is just not fair to judge the victim, who did not actively choose any steps leading to the outcome. Maybe in our ordinary praxis, we receive praise or blame for actions that are issued by a character we did not form actively, but in the presented manipulation cases, the disbalance seems obvious.

In addition, Vargas mentions that the condition of responsibility for one's character would be a sufficient but not necessary condition of moral responsibility in Mele's account[116].

**How to manipulate values**

The first objection to the examples of value manipulation could be that we don't know that value manipulation is possible even in principle. The possibility will depend on our model of thoughts. If we are representationalists, it seems plausible or at least imaginable that something such as values could be manipulated or replaced in us by some kind of advanced technology, because distinct objects in the world have distinct representations in our mind. However, according to some associationist theories of thought it makes much less sense. Associationist theories hold that an organism's causal history, agent's experience, is the main factor creating agential cognitive architecture by learning[117]. It means that our psychological structure basically is our history and the only way to replace it is to create some new history, new associations and new neural links which will become stronger than the old ones. It is related to the issue of learning to recognize moral

[116] Vargas, Building Better Beings, p.299

[117] What is Associationism? in Associationist Theories of Thought https://plato.stanford.edu/entries/associationist-thought/#WhaAss , last access: 5.8.2021, First published Tue Mar 17, 2015; substantive revision Wed Jun 24, 2020

considerations I mentioned in the previous chapter. I could only theorize that value manipulation would rather happen through a long process in isolation and radical change would be possible overnight only if the nefarious neurosurgeons have a machine with time-controlling technology or if we blast Sally in the space, where we prepare a lot of very strong and painful experiences for her to change her personality. For me, this is more fantasy than science fiction, so I would consider it impossible.

## 4.4 Where do their positions meet or differ?

In *Building Better Beings* Vargas mentions that he was previously inclined to think that Mele's account was an essentially historical one, because in his early works Mele uniformly appeals to a no-compulsion constraint. However, Vargas adds that threads of Mele's early account also suggested the following possibility: he could allow that possession or rare or exceptional agential powers might be sufficient for responsibility.

In context of agents globally manipulated by evil neurosurgeons, Vargas writes:

> "If Mele thinks that the paradigmatic or model agential structure required for responsibility is one in which an agent's relevant **desires are sheddable, then what I have been calling his "negative historical condition" is really the tracing condition of a [mixed] account**. On the other hand, if Mele does not think that there is any such requirement on the basic agential structure of responsibility, then his theory does count as a genuinely historical one. A case that is worth thinking about in this context is one where an agent has only sheddable desires—if Mele thinks that these structural properties (plus whatever other structural properties he thinks are required for responsible agency) are enough to make the agent a responsible

> agent, then on my way of thinking about these things, he is a mixed theorist. Given his stated account of compulsion, I am inclined to think that this is what he should say.[118]"

What does it mean? No-compulsion requirement suggests always looking into history but if we have the power to reflect and change our values at any time, we have a potential to undo any compulsion and the historical condition is losing importance.

Mele introduces an opposite term, *unsheddable values*. For him, values are understood as psychological states. If we have X as a value, we believe in it and desire it at the same time. -One may shed a value either by eradicating it or by significantly attenuating it. We can persuade ourselves that X is not that good or diminish our desire by reflection. If we extinguish such a desire completely, we have eradicated the value[119]. However, if values are unsheddable, they run deep and we cannot eradicate them easily.

> "An hour of careful reflection may be enough to disabuse oneself of an unwarranted belief that something is good, and an hour of sustained effort to weaken a desire by employing a self-control technique one has learned may suffice to weaken it significantly. But unsheddable values, as I understand them, are much more resistant to significant change than that.[120]"

They are supposed to be very firmly entrenched parts of the valuer's psyche.
"I will say that any agent who is stuck in this way with a value (during t) is practically unable to shed it (during t). Values that one is practically unable to shed

[118] "On the Importance of History for Responsible Agency," Philosophical Studies 127, no. 3 (2006): 376 n. 18.
[119] Mele, Manipulated agents, p. 42
[120] Mele, Manipulated agents, p. 44

may be termed practically unsheddable"[121]. He went on to say that "The notion of ability at work here is similar to one implicit in commonsense conceptions of irresistible desires"[122].

However, if an agent's desires are sheddable, there is no need to look into history, because agents can create their desires at any time.
Mele gave answer to Vargas' considerations by introducing Mabel[123]. Mabel is an agent with "the marvelous ability to produce in herself any conceptually possible system of values from moment to moment" and thus, "can undo the effects of value manipulation at any moment.[124]" By acknowledging the theoretical existence of such agents, Mele validates the search for internal conditions although his own approach is more practical and history dependent. The section above, describing Mele's classification from Manipulated Agents published in 2019, confirms this interpretation[125].

The interesting question here is whether there are any other presumptions in which Vargas and Mele differ and which make Mele focus way more on historical conditions or if it's just their respective research preferences. For sure, we can tell that Mele's argumentative approach is from positions of intuitions and interhuman praxis, whereas Vargas is trying to construct a full working concept of moral responsibility. I believe that these diverse approaches are valuable. Mele's work plays an important role in keeping the feet of those who are trying to construct a full theory on the ground. Trying to make a complete theory, such as Vargas's, is also an

[121] Alfred Mele, *Autonomous Agents: From Self-Control to Autonomy* (New York: Oxford University Press, 1995), p. 153
[122] ibid. p. 154
[123] first: Alfred Mele, "Moral Responsibility and History Revisited," Ethical Theory and Moral Practice 12, no. 5 (2009): 468., second in MA
[124] Mele, Manipulated agents, p. 15
[125] Vargas's book Building Better Beings, where he first mentioned the hypothesis that Mele's position could be consistent with conditional internalism, was published in 2013.

important approach, because only full theories will help us make sense of the robustness and complexity of the problem we are working with.

What motivates Vargas to pursue the internal condition even though agents like us ordinarily lack Mabel's marvelous power? Vargas concludes that both he and Mele are mixed theory proponents. However, he adds that they are on the different poles of a continuum and he, unlike Mele, thinks that conditions sufficient for moral responsibility can ordinarily be satisfied without requiring satisfaction from some historical condition[126]. He emphasizes capacities to recognize and self-govern in light of moral considerations, which are more readily had[127]. In Mele's account, however, the cases where historical condition doesn't matter are rare or even only theoretical.

### Methodological differences

Another point to make is that they differ in the amount of importance they are giving to intuitions about the manipulation cases. Vargas's overall strategy towards intuitions is that they should not necessarily shape our accounts. They may be erroneous and if our prescriptive theory of moral responsibility offers a better solution, we should abandon them[128].

Mele's arguments are often based on the fact that the majority of people have a strong intuition that an agent is not morally responsible in manipulation cases. Vargas agrees that it's important to pursue why people have those intuitions, but unlike Mele, he doesn't think that they show us any important clues, but rather a

[126] Vargas, Building Better Beings, p.297
[127] ibid. p.299
[128] Ibid. p.297

common understanding which can turn out to be totally misleading. He writes that we need a proper theory behind it.

I would say that Mele's approach is more statistical, exploring the current state of the art of moral responsibility practices. He relies on the assumption that folk's intuitions developed for a reason, even though the body of prevalent beliefs about moral responsibility may be a bit out of date.

Vargas' approach aims to go beyond folk's intuitions and provide a theory, which is not only explanative, but also forms our intuitions based on the definitions and concepts it offers. He supposes that the majority of intuitions could be misleading and it's a task of a philosopher to help people overcome biases by a proper theory.

To sum it up, according to Vargas we don't know if the intuitions are reliable or truth-guiding, even though they may be widespread. Mele thinks at least that we should not discard them since they may show us a valid approach alternative to theoretizing which should be included in our theory. He writes:

"I certainly am not suggesting that intuitions are the final word on philosophical matters. Sometimes we find that our intuitions clash with the intuitions of others. When that happens, there often is room for discussion and progress […] If we were to set intuitions aside entirely, we would be significantly reducing our resources […]."

## 4.5 Personal Identity

Radical transformations of values, like the one seen in *One Good Day*, invite questions about personal identity. This is a more serious objection present in more manipulation case discussions, but I will provide just Vargas' explanation:

"[Another] variable that affects how we think about these cases one's theory of personal identity over time. If one operates with a psychological conception of personal identity, where one's identity over time (or perhaps just one's continuity) is secured by overlapping psychological ties, then the more dramatic the manipulation the more likely it is that we will have disrupted identity or continuity conditions.[129]" In other words, depending on our conception of personal identity, identity may or may not survive the manipulation. If it does not survive, we will have to ascribe moral responsibility to two different persons: pre-manipulation and post-manipulation one.

## 4.6 Why historicism and why not?

Motivation for historicism could be fairness, securing that agents are excused from responsibility in situations over which they have no control in a non-coercive but still very radical sense. As I wrote in the introduction, some mild influence and nudging is acceptable manipulation. But does it make sense to hold a globally manipulated agent responsible? In global manipulation cases, it seems like they have no way of knowing about the manipulation nor a way to prevent it next time.

It seems to me that the distinction between external and internal conditions for moral responsibility is not that important. There may be practical benefits in looking for clues in the external environment and examining an agent's historical upbringing or traces of psychologists breaking into the house rather than trying to assess an agent's psychological structure. Yes, it is a good strategy of structuralists to say that moral responsibility depends on an agent's psychology at the time. At least conceptually. I hope it will be possible in the future to work with a model of agential psyche, but right now we would have a hard time even telling if the agent is mentally ill or not. It seems to be conceptually easy to say that a mentally ill person

[129] Ibid. p.279

or someone who lacks a particular psychological structure is not morally responsible but what if there's no such a thing as a psychological structure in the sense we presuppose it? Or at least, what if there is no way of assigning a particular content to a particular brain activity? In combination with unreliability of agent's self assessment we are out of methods at the moment.

It is already the prevailing opinion among scientists that the experiments do not prove what they aimed to prove and the monitored brain signal merely shows a readiness potential and attention. From the experiments that followed it was clear that the signal was the same whether subjects chose to flick or not.

Regardless of the elegance of the psychological structure hypothesis, it would be better to suppose only a behavioral model of the same thing saying that the agent is responsible if they would act the same way in a similar situation. Then we can distinguish the shades of moral responsibility - if an agent would act the same way in the same possible world, a similar world without manipulation or in the same situation at a different time. Why argue if a conceptual model is internal or external if it's only an abstract model?

# 5. Conclusions

## 5.1 The three approaches to moral responsibility of manipulated agents

Moral responsibility is complex. There seem to be many ways to look at it and Frankfurt's *Real Self account*, Mele's historicism and Vargas' *Reasons account* are approaching the problem from very different angles. Even the question "Is a manipulated agent morally responsible?" appears to have a different meaning for each of them. Frankfurt is interested in the attributability of the behavior to the agent's real self based on his current mental state, Mele addresses whether the agent is accountable in a fair, backward-looking way, and Vargas tackles the problem of the beneficial agency cultivation model. In that context, it is not surprising that they offer different answers to the question of the manipulated agent's responsibility: yes, no, maybe. It almost seems that the best we can do is to take the ambiguous term of moral responsibility as just an umbrella term for many different problems of human agency. We can then break the discussion down into sub-discussions about moral responsibility and the extent to which:
(1) I was truly choosing the action on my own without doubts;
(2) I was accountable for the action given the circumstances and history; and
(3) it would be beneficial for me to get the blame.

I believe that particular approaches can make more or less sense depending on the situation. After all, it looks like all the scenarios abstracted from who is making the judgement. The judge has some kind of a relationship to the manipulated agent and is prompted to judge by it. In a random encounter where we do not care about the agent's real self or fair accountability offering a snap judgement of blame is maybe the right kind of feedback and we may even be considered nice to care (because as Vargas wrote, blaming could be costly[130]). But if our goal is to cultivate fairness, whether from the position of a theorist, a code of conduct author, or a teacher forming the rules in a class, we should also consider historical factors. The *Real Self* point of view in turn plays a role in knowing the character of the agent - it is supposed to distinguish behavior issuing from someone's heart from loosely held courses of actions chosen due to random ideas. All the three approaches have their strengths and weaknesses and there are optimal situations to use each of them. Although the *Real Self account* seems to be less relevant for our ordinary moral practice, it is valid in relationships based on a deep analysis of agent's motivations (in psychological praxis or close relationships).

Moreover, by attributing moral responsibility, we assume all sorts of concepts which are not included in models of responsible agency such as the one of Vargas': a moral theory, social contract or at least expectations imposed on agents, decision-making process and the process of adoption of reasons which play a role in it, and a lot more. Some of the assumptions are made explicitly: abstracting from moral theories or taking moral responsibility as blameworthiness. But reactions to manipulation cases show that there are other factors implicit to our theories which have the final word in granting the judgement about moral responsibility. The open discussion about those implicit factors and assumptions is missing, e.g. the extent to which fairness plays a role for each philosopher. I would also consider the decision-making process to be the weak point of most of the theories. The way philosophers handle psychological concepts such as desires, values and reasons is alarming. Values are not interchangeable constants used in reflection for validating our desires as Mele suggests. Vargas' model of responsible agency has some details, but it still fails to capture manipulation. What sense does a model make of responsible agency if it

[130] Vargas, Building Better Beings, p.242

lacks enough detail to explain where manipulated agents fail to control their decisions?

Because of the missing model of decision-making process and missing understanding of how it could possibly be impaired by the influence of others, including manipulation among the negative conditions for moral responsibility could be a better approach.

It redirects the effort to examining the external traces that manipulation happened. Moral progress could be another reason to treat manipulated agents the same way as coerced agents. We can imagine that a more advanced society is one where we know what manipulation looks like and we go after manipulators to keep the balance, because manipulators are the agents who need to be changed, not the victims.

The argument against treating manipulation as a negative condition for moral responsibility is that emphasizing individual responsibility could be beneficial for society as a prevention against manipulation in general. Agents could learn the tricks manipulators use and navigate themselves out of dangerous situations without depending on institutional protection. But if something already happened, it seems reasonable to put the pressure mainly on the manipulator, not on the victim, to prevent the practice from happening again. What, if anything, is there that could be used as a basis for blaming the manipulated agent? In the case of misinformation, a lot of tools for verification of information and a number of analyses already exist. The agent could be blamed for ignoring them e.g. before sharing some propaganda, or (such as in the case of Mr Balda) commiting a crime based on that misinformation. At least this is definitely something a manipulated agent should be blamed for. However, it is hard to tell in global manipulation cases, where there is no common knowledge that the threat even exists, let alone the availability of tools to avert its consequences.

## 5.2 Further development

I have considered two main factors (or maybe two types) of control: character control and volitional control. The requirement that manipulation cases place on morally responsible agency models is that they should have a component where they may be impaired and as a result not fulfilling the control condition. Mele explains sufficiently that in case of radical reversals the character control is impaired due to lost access to a carefully built system of values, something actively chosen by the agent, and thus it is unfair to blame them for acts issued from values that were imposed on them. I believe that the pressure on Vargas to find space for something similar in his theory is justified. Distinguishing non-standard psychological structures created in agents under pressure seems crucial to model manipulated agents. But Vargas does not seem to perceive the need to add more detail to his model and explain how the reasons (or values) are created in it. He proudly claims to be a structuralist. Even though the bare fact of an agent's reason-responsiveness or lack thereof does not explain much regarding manipulation cases, if we do not state how the ability was acquired. Another plausible strategy for him could be to stick with his forward-looking interpretation of blame and claim that we do not need to exclude manipulated agents from moral responsibility, because in our social practice only the resulting benefits of blame matter. But he did not choose this approach either and is insisting on the claim that all reasons-responsive agents are morally responsible.

The threshold of moral responsibility for manipulated agents is extremely hard to find, because they appear to have all the capacities needed to be morally responsible for their decisions. Certain treatments seem to reduce the question of moral responsibility into a binary decision, where any degree of moral responsibility is treated the same.

But this convenient reduction glosses over the heart of the matter: A premeditated immoral act is different from the behavior of a manipulated agent who is merely guilty of negligence. Equating them only obscures our understanding of moral agency principles.

I would argue in favor of further research on agential psychological structure, but I am also sceptical that a single demarcation criterion for moral responsibility can be found. Exclusion then seems to be the practical thing to do. This leaves us with Mele's approach. Excluding manipulated agents from moral responsibility is a convention. It is also a practical solution, because while there is no way to examine psychological states at the time of the action, we can look for traces of manipulation the same way we investigate coercion.

It is already the case that recorded threats serve as evidence against the aggressor and exonerate the victim. The same way online media policies could be based on recording what content the manipulated agent was exposed to and limits could be set for what counts as brainwashing.

# Bibliography

BARNES, Eric Christian. Character control and historical moral responsibility. *Philosophical Studies* [online]. 2016,173(9), 2311-2331 [cit. 2020-10-31]. ISSN 0031-8116. Dostupné z: doi:10.1007/s11098-015-0610-2

COONS, Christian a WEBER, Michael, ed. *Manipulation: Theory and Practice*. Massachusetts: Oxford University Press, 2014. ISBN 978-0-19-933821-4.

CLARKE, Randolph. Agent causation and the problem of luck. *Pacific Philosophical Quarterly* [online]. 2005,86(3), 408-421 [cit. 2020-10-29]. ISSN 0279-0750. Dostupné z: doi:10.1111/j.1468-0114.2005.00234.x

CLARKE, Randolph. "Motivation and Agency by Alfred R. Mele", *Mind* 113 (2004), p. 565-569

CYR, Taylor W. Manipulation and constitutive luck. *Philosophical Studies* [online]. 2020,177(8), 2381-2394 [cit. 2020-10-31]. ISSN 0031-8116. Dostupné z: doi:10.1007/s11098-019-01315-y

CYR, Taylor W. Manipulation Arguments and Libertarian Accounts of Free Will. *Journal of the American Philosophical Association* [online]. 2020,6(1), 57-73 [cit. 2020-10-31]. ISSN 2053-4477. Dostupné z: doi:10.1017/apa.2019.31

DENNETT, Daniel C. *Freedom Evolves.* New York: Viking Books, 2003.

DORIS, John M. "Doing without (arguing about) desert", *Philosophical Studies volume 172* (2015), p. 2625-2634

FISCHER, John Martin a RAVIZZA, Mark. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998. ISBN 978-0-521-77579-3.

FISCHER, John Martin. Responsibility and Manipulation. *The Journal of Ethics* [online]. 2004,8(2), 145-177 [cit. 2020-10-31]. ISSN 1382-4554. Dostupné z: doi:10.1023/B:JOET.0000018773.97209.84

FRANKFURT, Harry G. *The Importance of What We Care About: Philosophical Essays*. Cambridge: Cambridge University Press, 1988.

FRANKFURT, Harry G. " *Freedom of the will and the concept of a person*". *Journal of Philosophy 68* (1971), p. 5-20

FRANKFURT, Harry. "Reply to John Martin Fischer. ". In: Sarah Buss and Lee Overton (eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt*, p. 27-28. The MIT Press, 2002

FRANKLIN, Christopher Evan. Plausibility, Manipulation, and Fischer and Ravizza. *The Southern Journal of Philosophy* [online]. 2006,44(2), 173-192 [cit. 2020-10-31]. ISSN 00384283. Dostupné z: doi:10.1111/j.2041-6962.2006.tb00097.x

HAJI, Ishtiyaque a CUYPERS, Stefaan E.. Libertarian Free Will and CNC Manipulation. *Dialectica* [online]. 2001,55(3), 221-239 [cit. 2020-10-29]. ISSN 00122017. Dostupné z: doi:10.1111/j.1746-8361.2001.tb00217.x

HAJI, Ishtiyaque a CUYPERS, Stefaan E.. Moral responsibility and the problem of manipulation reconsidered. *International Journal of Philosophical Studies* [online].

2004,12(4), 439-464 [cit. 2020-10-29]. ISSN 0967-2559. Dostupné z: doi:10.1080/0967255042000278076

HARLAND, Harry. "Beyond the Moral Influence Theory? A Critical Examination of Vargas's Agency Cultivation Model of Responsibility", *The Journal of Ethics volume 24* (2020), p. 401-425

HUTCHISON, Katrina -- MACKENZIE, Catriona -- OSHANA, Marina. *Social dimensions of moral responsibility*. New York: Oxford University Press, 2018.

KAISERMAN, Alex. 'More of a Cause': Recent Work on Degrees of Causation and Responsibility. *Philosophy Compass* [online]. 2018,13(7) [cit. 2020-10-31]. ISSN 17479991. Dostupné z: doi:10.1111/phc3.12498

KEARNS, Stephen, "Aborting the zygote argument", *Philosophical Studies* 160 (2012), p. 379-389

KHOURY, Andrew C. Synchronic and diachronic responsibility. *Philosophical Studies* [online]. 2013,165(3), 735-752 [cit. 2020-10-29]. ISSN 0031-8116. Dostupné z: doi:10.1007/s11098-012-9976-6

MACKENZIE, Catriona. "The Importance of Relational Autonomy and Capabilities for an Ethics of Vulnerability". In: Catriona Mackenzie, Wendy Rogers, and Susan Dodds (eds.), *Vulnerability: New Essays in Ethics and Feminist Philosophy, p. 33. New York: Oxford University Press, 2014.*

MCKENNA, Michael. Responsibility and Globally Manipulated Agents. *Philosophical Topics* [online]. 2004,32(1), 169-192 [cit. 2020-10-31]. ISSN 0276-2080. Dostupné z: doi:10.5840/philtopics2004321/222

MCKENNA, Michael. Manipulation Arguments, Basic Desert, and Moral Responsibility: Assessing Derk Pereboom's Free Will, Agency, and Meaning in Life. *Criminal Law and Philosophy* [online]. 2017,11(3), 575-589 [cit. 2020-10-31]. ISSN 1871-9791. Dostupné z: doi:10.1007/s11572-015-9388-8

MCKENNA, Michael. "A Hard-line Reply to Pereboom's Four-Case Manipulation Argument", *Philosophy and Phenomenological Research 77* (2008), p. 142-159

MELE, Alfred R. *Manipulated Agents: A Window to Moral Responsibility*. New York: Oxford University Press, 2019. ISBN 9780190927967.

MELE, Alfred R. *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press, 1995. ISBN 0-19-509454-9.

MELE, ALfred R. *Motivation and Agency*. Oxford: Oxford University Press, 2003

MELE, Alfred R. Moral responsibility and manipulation: on a novel argument against historicism. *Philosophical Studies* [online]. 2020,177(10), 3143-3154 [cit. 2020-10-29]. ISSN 0031-8116. Dostupné z: doi:10.1007/s11098-019-01363-4

MELE, Alfred R. "Moral responsibility for actions: epistemic and freedom conditions". *Philosophical Explorations 13 (2010), p. 101-111.*

MELE, Alfred R., "Manipulation, Moral Responsibility, and Bullet Biting", *The Journal of Ethics* 17 (2013), p. 167-184

MELE, Alfred R., *Free Will and Luck*. Oxford: Oxford University Press, 2006

MURRAY, Dylan a LOMBROZO, Tania. Effects of Manipulation on Attributions of Causation, Free Will, and Moral Responsibility. *Cognitive Science* [online]. 2017,41(2), 447-481 [cit. 2020-10-31]. ISSN 03640213. Dostupné z: doi:10.1111/cogs.12338

PEREBOOM, Derk. *Free Will, Agency and Meaning in Life*. New York: Oxford University Press, 2014. ISBN 978-0-19-877686-4.

STRAWSON, Peter. "Freedom and Resentment", *Proceedings of the British Academy* 48 (1962), p. 1-25

*The Journal of Ethics: Free Will and Moral Responsibility: Three Recent Views*. 1997-2016. Springer, 2000. ISSN 13824554.

VARGAS, Manuel. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press, 2013. ISBN 978-0-19-870936-7.

WALLER, Bruce N. *Against Moral Responsibility*. Massachusetts: The MIT Press, 2011. ISBN 978-0-262-01659-9.

WATSON, Gary. Free Action and Free Will. *Mind* [online]. 1987,XCVI(382), 145-172 [cit. 2020-10-31]. ISSN 0026-4423. Dostupné z: doi:10.1093/mind/XCVI.382.145

WILLEMSEN, Pascale -- NEWEN, Albert -- KASPAR, Kai."A new look at the attribution of moral responsibility: The underestimated relevance of social roles", *Philosophical Psychology 31* (2018), p. 595-608.