Freedom, Reason, and Holding Responsible

Ivo Dragoun

Philosophy PhD

University of Sussex

September 2019

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature ………………………………………………

University of Sussex

Ivo Dragoun

Philosophy PhD

Freedom, Reason, and Holding Responsible

Summary

In this thesis, I present a novel compatibilist solution to the problem of free will. The presented solution rests on three strategic pillars.

The first pillar: It is widely accepted that the justifiability of our common practice of holding others morally responsible is seriously undermined should it turn out that free will is impossible. This threat to the practice is what motivates the discussion of the problem of free will. However, the motivation is not, I will argue, theoretically innocent. It comes part and parcel with certain deeply misleading constraints. In Chapter 1, I challenge the widely accepted background motivation and the associated constraints. The insight generated by the discussion in Chapter 1 allows me to approach the problem of free will from a different angle.

The second pillar: The vast majority of philosophers agree that free will can be attributed only to agents that are, in some suitably robust sense, rational; that is, they see rationality as a necessary condition of attributability of free will. In Chapter 2, I formulate an original argument in support of this view. In Chapter 3, I present a thought experiment designed to show that, and how, rationality can be understood as not only a necessary condition but a condition that is, in fact, sufficient for the attributability. I conclude that acting rationally is all there is to acting freely.

The third pillar: The logic of my account of how rationality can be understood as a sufficient condition of the attributability of free will is such that it won`t help neutralize the threat that universal causal determinism poses to the justifiability of our practice of holding others morally responsible. This theoretical impotence will be perceived as a weakness of the account. The last two chapters of the thesis are meant to neutralize the perceived weakness.

It is generally, and mostly implicitly, assumed that once it has been established that we have free will, there won`t be any other principle reason that undermines the justifiability of our practice of holding others morally responsible. I challenge the assumption. I argue that there is another such principal reason in virtue of which the practice is irremediably unjustifiable. To be able to argue so, I first identify the kind of holding morally responsible that calls for the free will assumption. This is done in Chapter 4. In Chapter 5, I argue that the

kind of holding morally responsible that calls for the free will assumption is in principle unjustifiable due to the morally corrupted essence of man. I conclude that the theoretical impotence of my account of free will regarding the issue of the justifiability of our practice of holding others morally responsible cannot be taken as a weakness of the account because the practice is, for a separate reason, irremediably unjustifiable anyway.

In Appendix, I offer a separate argument for the claim that man`s moral standing is essentially corrupted. The argument relies on an interpretation of a passage from the New Testament and as such its appeal is somewhat limited. The target audience of the argument is a broadly Christian reader. Because of the limited appeal of the argument, it hasn`t been included in the main body of the thesis.

Table of Contents

Acknowledgement

Three things have made the completion of the thesis possible:

The patient and subtle guidance of my supervisor, Prof. Michael Morris. The unconditional support of my wife Baha. And the prodigious ability of my daughter Annika to make me laugh whenever I struggled most.

To all three I will stay greatly indebted.

Introduction

It's only by thinking even more crazily than philosophers do
that you can solve their problems.

Ludwig Wittgenstein[1]

The problem of free will is one of the most discussed problems in the history of philosophy. It is so because much about who we are as moral and socially interacting agents seems to turn on the answer provided.

Humans care about the justifications of their actions. Among the actions where they seem to care most about justifications is holding others morally responsible. The practice of holding others morally responsible is constitutive of our everyday social interactions. It is generally agreed that the moral justifiability of holding another responsible is severely compromised in cases where the person held responsible lacks free will. Thus, should it turn out that there is some issue of principle with free will and agency, then the socially constitutive practise of holding others responsible would be seriously undermined. And, it is a widely shared worry that, indeed, there is at least one such issue of principle with free will and agency.

The issue with free will and agency I have in mind here and will address in this thesis has the form of a dilemma constituted by two apparently and mutually contradictory assumptions:[2]

People possess free will.

The world is such that there are no alternative futures.

The assumptions are understood, at least on the face of it, as mutually contradictory because the possession of free will is taken to involve an availability of alternative future actions that an agent possessing free will is choosing from, which, at the same time, is something that is explicitly ruled out by the second assumption.

[1] Ludwig Wittgenstein, *Culture and Value*, ed. G.H. von Wright, trans. Peter Winch (Chicago: The University of Chicago Press, 1980) p. 75e.

[2] This is not to say that this is the only understanding of what constitutes the problem. Depending on their other theoretical commitments, some philosophers see the problem as being about the tension between the divine foreknowledge and human free will, or (if they are libertarians) about figuring out how an indeterminate causation could be embedded in autonomous agency, or they might simply take it to be about answering the question whether or not determinism is true.

The majority of philosophers working on the problem of free will as understood here are unwilling to `dissolve` the dilemma by giving up one of the assumptions. The first assumption is deeply grounded in the phenomenology of our everyday experience with our own agency. The second assumption is entailed in the thesis of universal causal determinism which is, in contemporary discussions, taken as foundational for the highly respected project of natural sciences to explain the world in terms of causes and laws of nature. Thus, giving up either of the assumptions would incur unacceptable theoretical costs. Most philosophers will, therefore, accept the assumptions and attempt to show how the apparent contradiction could be removed.[3] In this thesis I shall attempt to do the same.

The apparent contradiction between the two assumptions has been a very resilient one. None of the proposed resolutions seems to have been particularly successful, as evidenced by the undying interest of both philosophers and the general public in the problem. When a theoretical problem has resisted a solution for a long time, one can do worse than re-examine the conceptual framework within which the problem is typically tackled. Such a re-examination can, with a bit of luck, yield a conceptual framework that is more solution-friendly.

I engage in such a re-examination in Chapter 1. More specifically, I scrutinize an assumption that seems to be implicit (or explicit) in vast majority of the contemporary attempts to solve the problem of free will. The assumption could be put as follows: Any successful solution to the problem of free will has to be such that it neutralizes the threat that the second horn of the dilemma (*The world is such that there are no alternative futures*) poses regarding the justifiability of our practice of holding others (morally) responsible.

In philosophical discussions of the problem of free will, the assumption manifests itself through philosophers` using, in many contexts, the notions of *free will* and *moral responsibility* interchangeably. This interchangeability is often seen as innocuous because it is accepted that there is a robust conceptual dependency between the two notions. A proposed account of free will is, typically, tested in thought experiments that are designed to elicit intuitions regarding

[3] There are philosophers who reject the dilemma as they see no tension between determinism and alternative possibilities. Daniel Dennett, for instance, cleverly equivocates between ontological and epistemological readings of determinism to be able to claim that, for all practical purposes, there are options to choose from despite the truth of determinism. Another good example of a philosopher rejecting the dilemma is David Lewis who argues that the dilemma collapses once we disambiguate the notion of `could have done otherwise`. Once disambiguated, it can be seen that there is a robust sense in which one can be understood as having alternatives in a determined world. For details see Daniel Dennett, *Elbow room: the varieties of free will worth wanting* (Cambridge, Mass.: MIT Press: 1984), especially the last chapter, and David Lewis, `Are we free to break the laws?`, *Theoria*, 47 (1981), 113-121. It is worth noticing that although these philosophers formally reject the dilemma, it's not implausible to interpret them as providing an analysis which avoids it, precisely because they take the dilemma seriously; and if they are understood like that, they too can be seen as starting from the same position.

the justifiability of holding, in a particular scenario, a person morally responsible. The moral intuitions appealed to in those thought experiments are – for reasons discussed in Chapter 1 - extremely difficult (perhaps impossible) to satisfy.

Now, what if it turned out that, not only, we don`t have to approach the problem of free will constrained by the need to satisfy the moral intuitions, we even *shouldn`t* do so, as such a constrained approach is intrinsically deeply misleading? Surely, it is not unreasonable to expect that once this constraint is removed, new, hitherto neglected, theoretical avenues towards the solution of the problem of free will might open. Chapter 1 is dedicated to a removal of the constraint.

Chapters 2 and 3 capitalize on the theoretical gains generated by the discussion in the first chapter. Having disconnected the discussion of free will from moral theory allows me to focus *exclusively* on the conceptual architecture of *free will* and how it relates to the logic of causal determinism; that is, put simply: I don`t have to worry about answering the (relatively more difficult) question of how *moral responsibility* is possible in a causally determined world.

In Chapter 2, I argue towards the claim that an exercise of practical reasoning is a *necessary* condition of attributability of freedom to an agent. The claim, or a version of it, is not terribly controversial as the majority of philosophers working on freedom and agency will agree that only some suitably rational agency is describable as free. They will, however, agree so for reasons (that I will touch upon in Chapter 2) that I find unsatisfactory. Therefore, I offer an argument of my own towards the claim.

In Chapter 3, I argue that an exercise of practical reasoning is not merely a *necessary* but, in fact, a *sufficient* condition of attributability of freedom to an agent. A crucial role in the argument is played by an original thought experiment designed to show that whenever an action is *rational*, it is also *free*. The thought experiment explicitly assumes that the world is causally determined thus the conclusion of the argument constitutes a novel *compatibilist* solution to the problem of free will.

A conspicuous feature of the compatibilist solution to the problem of free will presented in Chapter 3 is its theoretical impotence regarding the problem of grounding our practice of holding others morally responsible in the causally determined world. I explain, in Chapter 1, why it is confused and misleading to approach the problem of free will under the assumption that a solution to the problem will be recognized as successful only if it does the grounding. Still, it will be, I suspect, seen by many as a weakness that the solution to the problem of free will proposed here is of no use to a moral theorist. Surely, most philosophers working on the problem of free will are motivated by the issue of the grounding. The last two chapters of the thesis are dedicated to neutralizing this apparent weakness.

As mentioned above, most philosophers working on the problem of free will are motivated by worries about grounding our practice of holding others morally responsible in the world that is fully causally determined. The motivation comes part and parcel with an implicit assumption that once it is shown how an agent can be free in such a causally determined world, there is no other serious theoretical obstacle to taking the practice of holding others morally responsible as safely grounded.[4] My strategy in Chapters 4 and 5 will be to show that the problem of free will is not the only serious threat to the justifiability of the practice. There is another serious threat, and this other threat cannot, I shall argue, be neutralized. The implication of this claim regarding the apparent weakness of my compatibilist solution to the problem of free will is this. My compatibilist solution won`t help with the grounding of the practice. The practise is, however, ungroundable anyway because of a separate threat that cannot be neutralized. This means that *all* solutions to the problem of free will are going to suffer from the same weakness. This particular weakness thus cannot be selectively pointed out as a relative disadvantage of my solution.

Most of Chapter 4 is devoted to disambiguating the notion of holding morally responsible. The idea behind this here is that there are different kinds of holding morally responsible and that not all of them call for the free will assumption, that is, require grounding in the assumption that we are free agents. Only after we have disambiguated the notion of holding morally responsible can we identify the kind of holding so that calls for the free will assumption. The path towards this identification will reveal that in our practice (of the kind that calls for the free will assumption) of holding others morally responsible, we are motivated by an urge to restore *fairness* disrupted by a wrongdoing. This motivation by *fairness* is, it will turn out, what lies behind philosophers` interest in the problem of free will. Philosophers working on the problem, typically and as mentioned above, see the problem of free will as being about explaining how moral responsibility is possible in the causally determined world. They see it so because they agree that it would be *unfair* to engage in holding others morally responsible in a world where free will is impossible.

The discovery that holding others morally responsible has essentially something to do with *fairness* and its restoration (and/or preservation) gives us a conceptual tool to examine whether the kind of holding morally responsible that calls for the free will assumption can, in fact, ever be *fair* even if it can be somehow shown that we are free agents. The conceptual tool will be deployed in Chapter 5 to argue that our practice (of the kind that calls for the free will assumption) of holding others morally responsible is, indeed, essentially and irremediably unfair.

[4] That is: to take it as, in principle, *justifiable*.

In Chapter 5, I make use of the uncontroversial claim that the appropriateness of blaming others for their wrongdoing depends, among other things, on the *moral standing* of the blamer. Against this background, I draw an analogy between the practice of blaming and that of holding morally responsible and claim – again, fairly uncontroversially – that the *fairness* of holding another morally responsible depends, among others, on the moral standing of the holder. It depends so in the following sense: an act of holding another morally responsible is not *fair* if the moral standing of the holder is compromised. I proceed to formulate an argument designed to convince the reader that as moral beings we are such that our moral standing is essentially and irremediably compromised and that, consequently, our practice (of the kind that calls for the free will assumption) of holding others morally responsible is essentially and irremediably *unfair*. I conclude that it is no weakness of my account of free action that it won`t ground the practice of holding others moral responsible because the practice is – for a separate reason – ungroundable anyway.

The idea that man`s moral standing is essentially compromised is – at least in the Christian tradition – an old one. In the Appendix, I offer an alternative argument for the idea. The argument relies, crucially, on a philosophical interpretation of a passage from the New Testament. Nowadays, the vast majority of philosophers will see religious texts as largely irrelevant in the context of solving philosophical problems. Thus a substantial part of the chapter is devoted to showing why at least those philosophers who are moral realists should take religious texts as relevant to their theoretical research in ethics. I formulate two sub-arguments to show that. The first sub-argument is rather controversial as in one of its steps I argue that moral realism implies the existence of God. The second sub-argument will be considerably more convincing as it doesn`t rely on any assumption that a moral realist should find particularly problematic.

After the two sub-arguments have been presented and discussed, I proceed to the central task of the Appendix, which is the extraction of philosophical reasons from a passage in the New Testament. I show how the passage can plausibly be understood as entailing that it is essential to all men that their moral standing is compromised. This fact then makes the practice (of the kind that calls for the free will assumption) of holding others morally responsible essentially and irremediably *unfair*.

Within the thesis, the alternative argument is presented in the Appendix because its dependency on a canonical Christian text limits its potential appeal to broadly Christian philosophers only.

The compatibilist picture of free agency presented in the thesis rests on three pillars. The first pillar gets erected in Chapter 1 and is about exposing the misleading conceptual dependency of the discussion of the problem of free will on moral theory. The exposition liberates the discussion and opens new avenues towards a solution of the problem. The second pillar gets erected in Chapters 2 and 3 and gives us the core compatibilist argument that allows us to see how free agency is possible in a causally determined world. Chapters 4 and 5 constitute the third pillar, a pillar whose role in the overall picture is to neutralize a strategic weakness that most philosophers will perceive the proposed solution to suffer from.

The three pillars are independent of each other. This means that accepting or rejecting any one of the pillars is inconsequential regarding accepting or rejecting the other ones.

In Appendix, I offer a separate argument for the claim that man`s moral standing is essentially corrupted. The argument relies on an interpretation of a passage from the New Testament and as such its appeal is somewhat limited. The target audience of the argument is a broadly Christian reader. Because of the limited appeal of the argument, it hasn`t been included in the main body of the thesis.

Chapter 1: Freedom and morality: severing the connection

1.1 Introduction

The central claim of the chapter is that the concept and discussion of free will/agency[5] can and should be separated from the concept and discussion of moral responsibility.

Ted Honderich claims that both compatibilists and incompatibilists typically understand the problem of free will/agency as a problem that can be settled linguistically:

> Both sides agree that what we have to do is just to see clearly, not get confused, get a good definition of the idea we all share, not get led astray by other philosophers with a doctrinal ax to grind…The question is importantly a linguistic one. What we have to do is analyse \`free\` in ordinary English and similar words in other languages.[6]

I agree with Honderich that the problem of free will/agency is a linguistic (or conceptual) one. Thus, it would be philosophically relevant if we were to discover that our grasp of the key concept(s) is somehow incorrect or misleading. In this chapter I will, first, argue that connecting the concept of free will/agency with that of moral responsibility in the way it is often done in contemporary literature is unwarranted and as such should be severed and, second, show how this unwarranted connection necessarily leads to some truly puzzling issues that could be avoided once the relationship between free will/agency and moral responsibility is clarified and, ultimately, severed. The first step provides *a-priori* considerations. The second one provides *a-posteriori* considerations that should motivate us to sever the connection.

1.2 The first step

There seems to be a general and rarely disputed consensus that the concept of free will is fundamentally connected with the issue of moral responsibility. Claims and assumptions (both explicit and implicit) of the following kind are very common in the relevant literature:

> \`Free will is thought of as the capacity to be genuinely responsible for actions and genuinely deserving of praise or blame for the choices that are made.[7]

[5] I shall often use \`free will/agency\` or \`freedom in agency\` instead of just \`free will\` throughout the thesis. The reason for using this somewhat cumbersome replacement is the following: The notion of (free) *will* is a power notion; i.e. a notion that constrains our conceptual imagination and locks our theorizing about the problem of free will within the assumption that being free in the causally deterministic world must involve some kind of *power* that the agent possesses. I see this as an unwarranted constraint which, possibly, hinders progress in solving the problem. A neutral and not misleading alternative to (free) \`will\` is (free) \`agency\`. At the same time, I don\`t want the reader to forget for a moment that we are discussing what is known as the problem of free will. The expression \`free will/agency\` or \`freedom in agency\` is meant to capture both points.

[6] Ted Honderich, *How Free Are You?* (Oxford: OUP, 1993), p.101.

[7] Janet Radcliffe Richards, *Human Nature after Darwin* (New York: Routledge, 2000), p.136.

> `Free will is just the capacity that gives persons the relevant sort of control required for morally responsible agency`.[8]

> `[A] central aspect of the free-will problem is the problem of explaining how it can even make sense to hold people responsible for what they do. To this extent, in this context, the most important sense of `free` is perhaps just this: someone is free insofar as they satisfy whatever conditions needs to be satisfied for this to make sense with respect to them`.[9]

> `If we do know that moral responsibility exists, then we should have no doubt about whether we have good reason to believe we have free will…It is as adequate a defence of the free-will thesis as has ever been given for any philosophical position to say, ‘ Without free will, we should never be morally responsible for anything, and we are sometimes morally responsible``.[10]

However, I claim, tying free will/agency with moral responsibility in the spirit of the quotes above is ambiguous: in one sense it is warranted and in the other it isn`t. It is correct to take the concept of moral responsibility as – in a certain robust sense – depending on (and in this way tied to) the concept of free will/agency but it is incorrect to take the concept of free will/agency as depending on the concept of moral responsibility. In other words: there is a relationship of dependency between the concept of moral responsibility and the concept of free will/agency. The dependency, however, is not mutual but runs in one direction only. Often, this point goes unnoticed in the contemporary literature on free will, which, I claim, leads to a deep confusion that seriously affects the whole discourse.[11]

Before I proceed to discussing the deep confusion, a caveat needs to be lodged: I don`t wish to claim that *all* philosophers working on the problem of free will conflate or tie the notions of freedom and moral responsibility together in the way that will be criticised below. There are some who not only don`t conflate the notions in any illegitimate way but who explicitly criticize any such conflation. Kadri Vihvelin, for instance, says about the assumption that `to be a free agent is, […], to have what it takes to be morally responsible for our actions`, that `this way of approaching the free will/determinism problem is a mistake`.[12] And some other philosophers, although not explicit about the conflation, stay clearly consistent in their approaching the problem as independent of moral theory. A good example is David Lewis`s

[8] Michael McKenna, `Ultimacy and Sweet Jane` in Nick Trakakis and Daniel Cohen (eds), *Essays on free will and moral responsibility* (New Castle upon Tyne, UK: Cambridge Scholars Publishing, 2008), pp.187-188.

[9] James Lenman, `Compatibilism and contractualism: The possibility of moral responsibility`, *Ethics*, 117, no.1 (2006), pp.7–31, (p.8).

[10] Peter van Inwagen, *An essay on free will* (Oxford: Clarendon Press, 1983), p.209.

[11] Philip Pettit notices something very similar to this point of mine here – in his *A Theory of Freedom* (Oxford: Blackwell, 2001), p.32 – `[W]e ordinary folk think of freedom in purely functional terms. We conceive of it as that capacity, whatever it involves in itself, in virtue of which an agent is fit to be held responsible, satisfying the various constraints that that involves. Freedom is identified for us, then by reference to the function it plays in making agents fit to be held responsible and its essential character – what it is in itself – is left in the dark. This functional characterization immediately opens up the question as to what the capacity involves in itself. And that is the question that theories of freedom, as I think of them, should address`.

[12] Kadri Vihvelin, `How to Think about the Free Will/Determinism Problem` in Joseph Keim Campbell, Michael O`Rourke and Matthew H. Slater (eds.), *Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science* (Cambridge, Massachusetts: MIT Press, 2011), pp. 313-340, (p.315).

compatibilism which is grounded solely in his ingenious treatment of the semantic behaviour of the phrase `could have done otherwise`.[13] Despite the existence of a minority of philosophers who don`t tie the notions of freedom and moral responsibility together in some illegitimate way, it seems to be a fact that most of the contemporary discourse on the problem of free will is guilty of discussing the problem as constrained by conceptual ties to moral responsibility. Kadri Vihvelin agrees:

> In the contemporary literature, the free will/determinism problem is almost invariably viewed through the lens of moral responsibility. It is, of course, widely agreed that having free will is a necessary condition of being a morally responsible agent. But most contemporary discussions of the free will/determinism problem forge a much stronger link between questions about moral responsibility and questions about free will.[14]

Thus, there remains the task of showing what exactly is wrong with viewing the problem of free will `through the lens of moral responsibility`.

So, what is meant by saying, above, that the concept of moral responsibility depends (in some sense) on the concept of free will/agency but not vice versa? It seems to be a fairly uncontroversial fact about moral responsibility that it can be justifiably attributed to an agent only if the agent has free will/agency. There is a strong modal (perhaps not necessary but close to it) connection between moral responsibility and free will such that the former is attributable only if the former too is. The modal strength of the connection becomes quite obvious once we try to inspect the concept of moral responsibility as theoretically projected in a possible world where free will as a concept doesn`t exist. In such a world the concept of moral responsibility becomes rather unintelligible. How or in what sense do you hold those around you morally responsible for their actions if it never even occurs to you that they could have chosen otherwise? Is it even as much as conceivable that you would possess a concept of moral responsibility in such a world? I am unable to see how one could happen to arrive at a possession of such a concept at all. In a world where no one possesses the concept of free will/agency, no one possesses a concept of moral responsibility either. This is the sense in which the concept of moral responsibility depends fundamentally on the concept of free will/agency. As noted above this is not a controversial claim. I believe that considerations running along similar lines would explain why philosophers – in many contexts - use the two concepts interchangeably when theorizing on free will and moral responsibility.

1.2.1 *The perfect people scenario*

While it seems to be clear that the concept of moral responsibility depends on the concept of free will/agency, it is far from clear that the latter depends on the former in some such similar

[13] David Lewis, `Are we free to break the laws?`, *Theoria* 47 (1981), pp.113-121.

[14] Kadri Vihvelin, `How to Think about the Free Will/Determinism Problem`, p.314.

way; i.e., it is not – not *immediately,* at least – obvious that we cannot have a grasp of a rich and coherent concept of free will/agency in a possible world in which moral distinctions do not exist and in which, consequently, there is no concept of moral responsibility. Consider a hypothetical case of a community of morally and spiritually *perfect people*. Most of the descriptive details of such a community will be, more or less, irrelevant and can vary considerably. However, two things need to stay fixed across varying descriptions: (i) the members of the community – *the perfect people* – must lack any intentions that could be described as morally wrong or evil, and (ii) *the perfect people* must be beings that are rational in the way that is – for all or most practical purposes – indistinguishable from our rationality. Now, the absence of morally charged intentions means that in the community there is a general lack of awareness of any moral properties.[15] Consequently there is no moral language to describe those properties and no discourse on moral responsibility. Yet there seems to be no difficulty in conceiving that some of the *perfect people* - those with philosophical leaning perhaps - might be intrigued by and discuss the theoretical tension between the apparent nomological character of the world and the intimate acquaintance they have with willing freely. They might wonder whether perhaps that intimate acquaintance is just an illusion of a sort and generally find the whole issue extremely intriguing despite their lack of the concept of moral responsibility. They might – perfectly meaningfully - ask: are the world and humans such that humans are capable of acting in the world in some such way that it would be correct to describe them as free agents?

Although I find the *perfect people* scenario well within our normal capacity to conceive hypothetically, others might find it helpful here to be reminded of a well-known and entertaining precedent. The main character of Jonathan Swifts` famous prose satire *Gulliver`s Travels* Lemuel Gulliver ends up – after mutiny against him – left and abandoned by his crew on the first piece of land they come across. Here Gulliver comes upon Houyhnhnms, a race of talking horses. The horses `are purely rational beings that face no conflict between moral principles and impulses`.[16] The absence of such a conflict between moral principles and impulses gets reflected in their language. Houyhnhnms `have no Word in their Language to express any thing that is evil`.[17] I take it that having no concept of evil entails having no concept of good either. The fact that millions of readers of *Gulliver`s Travels* seem to have no

[15] The relation between the lack of morally charged intentions and the lack of awareness of moral properties is quite straightforward. Intentions are empirically accessible only via the first-person perspective. Once accessed, they can be projected onto other people and their actions. If a subject doesn`t ever detect – within her awareness – any morally charged intentions she won`t be able to see or detect those in others either and even the most harmful actions of other agents will be interpreted as (merely) regrettable accidents. Whatever moral properties the world might have they remain completely invisible for a subject that lacks morally charged intentions.

[16] Jonathan Swift, *The Writings of Jonathan Swift,* Robert A. Greenberg and William B. Piper (eds) (New York: W. W. Norton, 1973), p.240.

[17] Ibid.

difficulties make-believing Swift`s description of the character of Houyhnhnms as creatures that are highly rational and at the same time lacking awareness of a moral dimension of the world shows that something very close to our *perfect people* scenario is easily conceivable.

Now, in the case of the *perfect people*, I have claimed that there is nothing that could prevent them from having a robust concept of free will.[18] Here the analogy with Houyhnhnms helps only indirectly. Gulliver`s account of his encounter and experience with Houyhnhnms doesn`t mention what they might have thought about free will or whether they even as much as talked about it. However, he mentions that they liked to meet their friends to `cultivate reason`[19] and among the subjects they liked to talk about was `the visible Operations of Nature`.[20] Now, in Houyhnhnms we encounter creatures whose rationality we would recognize as more or less indistinguishable from that of ours and who show – again, similarly to us – a serious theoretical interest in the external world. I find it hard to see what, in principle, could prevent these rational creatures from noticing the nomological nature of the world and wonder whether or not their own actions are similarly necessitated as the events out there in the objective world. Clearly, this sort of wondering would require a grasp of suitable concepts

[18] It could be objected that the perfect people would not be able to have an intimate acquaintance with free willing because such an acquaintance requires an awareness of *conflicting* alternatives or desires, such as killing someone versus letting her live, where the nature of the conflict is essentially *moral* (or reducible to the moral). The idea here is that not any set of alternatives (desires) as a focus of awareness suffices for formation of the concept of free will/agency. There must be a conflict of certain strength among the alternatives or desires for it to sustain the awareness of choice and a formation of the concept of free will. At this point it is insisted that only *moral* alternatives/desires are robust enough to sustain the awareness of choice and the resulting formation of the concept of free will. Or, in other words, only *moral* alternatives (desires) are – in this respect - *constitutively* conflicting. The perfect people, however, lack such conflicting desires and it is this lack of conflicting desires that prevents them to form a notion of free willing. No conflicting desires means no awareness of conflicting alternatives. No awareness of conflicting alternatives means no awareness of choice and no awareness of choice means no intimate acquaintance with free willing. This objection amounts to saying that the grasp of the concept of (or acquaintance with) free will crucially depends on the awareness of moral distinctions and as such the concept cannot be separated from the discussion and the concept of moral responsibility. This, of course, contradicts my claim that the concept of free will/agency is independent of the concept of moral responsibility. There is some truth to the logic of the objection. It seems to be quite plausible to claim that not just any conflict between desires is strong or robust enough to sustain the awareness of choice (and a free choosing between them). The conflict must be *painful* enough to become conscious and moral conflicts are a very suitable (and only) candidate for a painful *enough* conflict. However, for the objection to take off it won`t suffice to identify moral conflicts as painful enough to generate the awareness of choice and need for the exercise of free will, it needs to insist that there are no (irreducibly) non-moral alternatives or desires that are sufficiently painful to sustain the formation of the concept of free will. And that seems rather implausible. Consider the following: Your wife and a daughter are drowning 100 meters away from each other. You find yourselves sitting in a boat exactly in the middle of the distance that separates them. It is highly probable that you will be able to – if at all – save only one of them. You are aware of this and you face a painful dilemma of having to decide which of the two you will try to save. This decision and choice is not of a moral kind – and cannot be reduced to one - yet it exemplifies the kind of a decision making process that involves choosing between alternatives that are clearly painful or conflicting enough to sustain the formation of the concept of free will/agency.

[19] Jonathan Swift, *The Writings of Jonathan Swift*, p.233.

[20] Ibid., p.242.

(and distinctions) such as *agency, causation, free will,* etc.[21] And again, I can`t see what could possibly – and in principle - prevent Houyhnhnms from engaging in subtle and sophisticated discussions about the possibility of free will/agency in a causally determined world despite them lacking any moral concepts.

1.2.2 Stipulating the connection?

My claim that the notion of free will/agency is conceptually independent of the notion of moral responsibility might be conceded at this point. However, this concession could be seen as a harmless one. In reply, something like the following could be suggested here: it is true, strictly speaking, that the concept of free will/agency is independent of the concept of moral responsibility in the sense explained above. We can accept this claim of independence and yet insist that it is a perfectly legitimate philosophical choice to focus on a kind of freedom that will have grounding our attributions of moral responsibility as its essential function. It might be, at the same time, admitted that there are other kinds of freedom worthy of theoretical investigation but the one that is essentially connected with justifying moral responsibility is by far the most interesting and important one. Richard Double sums up this point in the following way:

> Although we may elect to care about anything we like under the rubric of free will, perhaps the majority of philosophers who have written about free will have believed that justifying moral responsibility (including the expression of reactive attitudes, blame, punishment) is `the prize` that constitutes the point of caring about free will.[22]

In other words, it is, ultimately, a legitimate choice to tie the two concepts together in some such way that will serve a chosen theoretical goal. More specifically, there seems to be no reason why we can`t simply stipulate that the concept of free will/agency we are trying to provide an account of is to be such that it will ground attributions of moral responsibility; i.e. there seems to be no reason why we cannot stipulate that `free will/agency` *depends* in this sense on the concept of moral responsibility. And indeed, some such stipulation seems to be – explicitly or implicitly – taking place in the four passages quoted above (p. 2).

[21] Someone might wonder to what extent these concepts of the *perfect people* would overlap with those that we associate with the expressions above (i.e. *agency, causation, free will*) when in the community of the *perfect people* these concepts are not – unlike in our world – related to moral concepts. In other words, it could be objected that the *perfect people`s* concepts of agency, causation, free will are not and cannot be sufficiently similar to our concepts exactly because they are embedded in a different semantic web. My answer to the objection is part of the central claim of this chapter. I argue that the concept of free will/agency is conceptually independent from the concept of moral responsibility (although I agree that the concept of moral responsibility conceptually entails the concept of free will/agency) which, in the context of the objection, amounts to claiming two things: (i) some conceptual relations can be shown as contingent (or conceptually non-constitutive) and (ii) the relation between the concepts of free will/agency and moral responsibility is – with respect to the concept of free will/agency – of exactly such a contingent (or non-constitutive) nature.

[22] Richard Double, *Metaphilosophy and Free Will* (Oxford: Oxford University Press, 1996), p.11.

I believe there are reasons to suspect that such a stipulation - or choosing a kind of freedom one wishes to theorize about - is rather less innocent than it, *prima facie*, appears to be. Consider the following analogy:

We possess a concept of water and a concept of plant. We desire to have clear understanding of both concepts. We have empirically investigated the referents of both concepts. We have long known that plants require water for their existence. This physiological dependence of plants on water imposes certain explanatory requirements on our account of the concept of plant. Specifically, we will feel a theoretical urge to explain what about their essence[23] makes their existence depend on water. Our theoretical account of plants (and with it our conceptual grasp of them) will not be considered satisfactory if it doesn`t explain their physiological dependence on water. Our experiential knowledge of this physiological dependence of plants on water will direct our scientific research regarding the essence of plants in a particular way. The outcome of the research will – as it typically does – shape the conceptual content associated with *plant*.

What we have here is a sort of theoretical dependence of the notion of plant on the notion of water that is analogous to the dependence of *moral responsibility* on *free will*.

Crucially, this dependence runs in one direction only. The concept of water seems to be given fully to us to be grasped without any need for us to have any acquaintance with plants – that is, without possessing a concept of a plant[24]. Now, what happens if we stipulate – as is implicitly or explicitly done with the concepts of free will/agency and moral responsibility - that the dependence runs in *both* directions? That is, what happens if we adopt a requirement that we search for a theoretical account of water such that it is seriously (and in some sense essentially) incomplete if it doesn`t explain how water supports the existence of plants?

Adopting such a requirement would have absurd consequences. First, it would imply that scientists in the possible worlds where there are no plants couldn`t have a full grasp of the

[23] I am using the expression *essence* very loosely here, meaning something like *a constitution*. Thus, a scientist interested in the *essence* of a plant in the context of its physiological dependence on water can be said to be interested in what in the *constitution* of a plant makes its biological functions depend on water. There is no contentious philosophical commitment involved here.

[24] There is an issue here that needs to be briefly mentioned. In our world, a competent speaker will know that one of the things water does is that it sustains plants. If a plant owned by such a competent speaker dies after being `watered` with a liquid (presumed to be water), the competent speaker might suspect that she has *misidentified* the liquid. This suggests that, in our world, the property of sustaining plants is one of the crucial properties when it comes to *identifying* a liquid as water. Does this then imply that there is a conceptual dependency of *water* on *plant*? No, it doesn`t. What this implies is that, in our world, sustaining plants is one of the *identifying* properties of water. However, this particular property is just a relational property of water and as such cannot be understood as constitutive of water.

concept of water. That is, clearly, an extremely implausible implication.[25] The second implication is perhaps even more absurd: a countless number of possible worlds can be conceived of in which various kinds of entities depend for their existence on the existence of water in those worlds. In the actual world, we may never learn of the existence and nature of those various other water-dependent entities. If, however, our confusion regarding the conceptual dependence of *water* on whatever might, for its existence, depend on its referent (i.e. on water), imposes on us a requirement to theoretically account for this existential dependence - or else our conceptual grasp would have to be deemed seriously incomplete - we will find ourselves in a situation when, in the actual world, a robust grasp of the concept of water is simply impossible *in principle*. Once the *arbitrary* stipulations of conceptual dependence of the kind discussed above are allowed, we end up losing our grasp of the concept of water.[26]

Perhaps we could ignore these absurd consequences feeling no empathy with a scientist trying to understand water in a gloomy world with no plants and stay similarly unperturbed by those purely abstract worlds full of unknown and inaccessible entities that need water for their existence. Those are not *our* worlds and we have no epistemological responsibilities towards them. Still, even if we focus exclusively on theorizing about water as it is given to us in our actual world it must, I believe, feel to us a rather arbitrary and irrational requirement to search for an account of water such that it – in virtue of identifying a specific structural property of water – explains how it is able to sustain the physiological functions of plants. In other words, such an explanatory requirement commits us to look for an intrinsic property of water that we do not have any independent reason to expect to exist; and the default position should be one of *not* expecting it to exist.

At this point, several questions come to mind. Should it much surprise us if all attempts to formulate a successful account of water under this requirement keep failing? Would it not be appropriate to describe theorizing about water constrained by such a requirement as seriously misguided? Would it not be natural to expect theorizing misguided in such a way to reach an impasse sooner or later? I claim that the theorizing about free will is analogously

[25] I am trying to think what could be objected here. Perhaps, if the distinction between the *real* and *Cambridge* properties is denied than the relational (Cambridge) property that water possesses in virtue of being related to plants in our world could be seen as ineliminable from the concept of water. Then we could deny that the scientists in those possible worlds where there are no plants could ever have a full grasp of the concept of water. However, saving one absurd view (the view that scientist in the plant-less worlds cannot have a grasp of the concept of water) by introducing a similarly absurd assumption (the assumption that the *real* and *relational* properties can be conflated) is a strategy that merely multiplies absurdities.

[26] The reader will notice that permitting arbitrary stipulations of this kind would similarly affect our grasp of perhaps *any* concept we possess. We can conceive of infinite numbers of possible world with entities that in some robust sense depend on entities that our concepts refer to.

misguided by its stipulated conceptual dependence on the notion of moral responsibility and has, as a result of this misguided stipulation, reached an impasse.

Some might find the discussion of the *water-plant* dependency unconvincing. They could object that the conceptual dependence of *plant* on *water* derives from a physiological dependence which is a fact that seriously affects the plausibility of the analogy with the *free will/agency-moral responsibility* dependency. The latter dependence is of a different kind: it doesn`t derive from a physiological relation but is purely abstract. As such it is not subject to any external constraints. This objection is not too serious. Even purely abstract concepts can be shown to stand in a relationship of dependency that runs in one direction only. Consider the relation between the concept of sin and the concept of absolution. It is immediately obvious that *absolution* depends on *sin* as any absolution is always an absolution of a sin. There is, however, nothing similarly - that is, *immediately* - obvious once the reversed direction of conceptual dependence is inspected. It is hard to see how one`s grasp of the concept of sin and one`s competent usage of it could be understood as seriously affected or constrained in scenarios in which one has no grasp of the concept of absolution. The conceptual dependence seems to run in one direction only. *Absolution* depends on *sin* but not vice versa.

Now, let`s have a quick look at what we could expect happening once the dependence is stipulated. For instance, the following dependence of *sin* on *absolution* could be stipulated: something is a sin only if open[27] to absolution. What happens if we put the concept of sin constrained by such a stipulation to the test. It is agreed that the concept of absolution is absent from the world of the Old Testament scriptures[28] while a *sin* and related issues is one of its primary focuses. Shall we, perhaps, conclude that the Old Testament authors are not talking about sin after all? They don`t have a concept of absolution and therefore they cannot – once it is agreed that having a concept of sin requires having a concept of absolution – have a concept of a sin. But then, what do the authors refer to when using the word `sin`?[29]

Our puzzlement will get considerably stronger once we enter the conceptual world of the New Testament. The New Testament introduces Christ – the God of absolution – and ties the notion of sin with that of absolution in a way that conforms to the conceptual stipulation. This, however, comes at a price. The worlds of the Old and New Testaments are deeply intertwined. The authors of the New Testament had internalized the conceptual paradigm of

[27] To be open to absolution doesn`t mean that the absolution must be granted. It merely means that it is considered its target. Thus, a sin is open to absolution even if too serious to be granted any absolution.

[28] This is not to say that the Old Testament doesn`t refer to the forgiveness of sins. See, for instance, Leviticus 19:21-22. However, the concept of absolution goes way beyond the concept of (mere) forgiveness. Forgiveness restores the psychological and/or spiritual relation between the sinner and the one affected by the sin (often the God). Absolution undoes the sin itself.

[29] In Biblical Hebrew, the most common expression for sin is *hata*. Several other expressions for sin can be found in the Old Testament scriptures. Each of them comes with a specific connotation: *pesha* is a sin done out of rebelliousness; *aveira* is a sin as a transgression and *avone* is a sin as moral failing.

the Old Testament scriptures. The New Testament contains at least 302 direct quotes from and 493 allusions to the Old Testament.[30] There is a clear conceptual continuity in the usage of *sin* between the Old Testament and the New Testament. This continuity gets severed by the introduction of the stipulation and leads to serious exegetic problems. The problems might, perhaps, be solved but the solution will – highly likely – lack in explanatory simplicity compared to simply taking *sin* to be a concept that is independent of the concept of absolution.

There is a potential misunderstanding I would like to draw the reader`s attention to here. The *perfect people* scenario might be understood as an attempt to show that *necessarily* the concept of moral responsibility cannot be constitutively involved in the concept of free will/agency. Understanding the scenario in this way would be a mistake. It doesn`t follow from the scenario, and it is not what I need to claim as part of my argument. The perfect people have a grasp of *free will/agency* without having a grasp of moral concepts. Similarly, people in the 18th century had quite a robust grasp of *water* without having any idea about its chemical composition, i.e. without knowing that water = H2O. Later, in the 19th century, it turned out that being H2O is essential to being water. Analogously, it might turn out that a capacity to ground attributions of moral responsibility is an essential component of any coherent concept of free will/agency. Thus, the *perfect people* scenario must be understood as showing something weaker: as showing that the concept of moral responsibility is not *obviously* or *transparently* involved in the concept of free will/agency. It might, one day, turn out to be so but only as a result of a philosophical argument. And, as far as I am aware, such an argument has yet to be formulated.

Let me go back to the *sin – absolution* example to demonstrate what has just been said. Simple scenarios can be conceived of that will show that we can have a robust grasp of the concept of sin even without having any grasp of the concept of absolution. Now compare it with the complexities of an account attempting to show how *sin* - after a deeper conceptual investigation – turns out to entail *absolution*. Something like the following, for instance, could be suggested: the concept of sin is intelligible only in the context of a religious text. Religious texts presuppose a transcendental being, i.e. a god. God is an omnipotent being. That is, for God, there is nothing he cannot undo; whatever there is, can be made go away by Him. This applies to sins and their existence as well. It follows then that the concept of sin necessarily entails the possibility of its absolution. The two cannot be separated as without absolution the concept of sin lacks coherence.

Although I don`t find this particular line of argument very plausible, I cannot rule out that a version of it – or a different argument altogether – could turn out to be successful, committing

[30] The source for the figures given here is: <https://www.blueletterbible.org/study/misc/quotes.cfm> [accessed 12 November 2017].

us to take the concept of sin as entailing the concept of absolution. However, and that is crucial for my purpose here, this entailment can only be arrived at as a result of a philosophical argument and not *assumed* or *stipulated*. The concept of absolution is not given in the concept of sin in a *transparent* way, which is a fact that needs to be kept in mind when commencing any conceptual investigation of *sin*. Analogously, the concept of free will/agency does not – *transparently* – entail the concept of moral responsibility, which means that any conceptual investigation of the concept of free will/agency must start free of any commitment to conceptual constraints associated with *moral responsibility*.

1.2.3 Concerns about conceptual analysis

There is an issue here that needs to be addressed. The argument behind my claim that the concept of free will/agency is independent – and should be discussed as such – from the concept (and related issues) of moral responsibility relies heavily on conceptual analysis. Conceptual analysis used to be the undisputed *via regia* of philosophizing. However, ever since Quine`s attack on the analytic-synthetic distinction, conceptual analysis has at best been looked upon with suspicion or at worst outright dismissed as epistemically worthless. Now, how worried should we be about the viability of an argument that relies heavily on conceptual analysis, i.e. on a way of philosophizing that has fallen into considerable disrepute?

In what follows, I will briefly explain why I think there is no need to worry much. Let me first give you a sketch of the position of the critics. Conceptual analysis as understood by its critics is a *theoretical* activity that, (i), with respect to its goal, seeks to identify the necessary and/or sufficient conditions for applicability of a concept across possible worlds and, (ii), regarding its method, crucially relies on (modal) intuitions. Both (i) and (ii) are taken as hopelessly untenable by the critics of conceptual analysis. They argue – successfully, I believe – that there are no necessary and/or sufficient conditions that are both epistemically accessible in a priori fashion[31] and that could be used for a reliable identification of concepts across hypothetical scenarios. Conceptual analysis is epistemically worthless simply because `[t]here are no conceptual truths`[32] to be discovered. Regarding the prominent role intuition plays in conceptual analysis, the critics will, typically, refer to a considerable body of scientific research that shows how hopelessly unreliable our intuition is. This attack on intuition complements the attack on the very possibility of the existence of conceptual truths and is understood – by the critics - as fatal for the prospects of conceptual analysis: there are no

[31] The focus of the criticism is, specifically, on the implicit assumption that there are necessary/sufficient truths that can be discovered in the act of *a priori* analysis. Such criticism is perfectly compatible with endorsing Kripke`s *aposteriori* necessities.

[32] Timothy Williamson, `Conceptual Truth`, *Aristotelian Society Supplementary Volume*, 80, no. 1 (2006), pp.1-41 (p.39). In this paper, Williamson argues, convincingly, that even the most elementary conceptual/logical truths such as `Every vixen is a vixen` are not immune to the possibility of revision.

conceptual truths to be discovered from the philosopher`s armchair and even if there were some such truths they could not be reliably recovered by intuition as it is a fundamentally flawed faculty.

In response, let me first deal with the attack on intuition. As mentioned above, the critics back their claims about hopeless unreliability of intuition by a considerable body of scientific evidence. However, a closer inspection of this scientific evidence reveals that what scientists typically call `intuition` - and what they thus focus on in their experiments – is different from the intuition that the advocates of conceptual analysis have in mind. George Bealer calls the intuition that the advocates have in mind *rational* intuition.[33] The scientific evidence referred to by the critics is irrelevant as it doesn`t target the *rational* intuition, i.e. the kind of intuition employed in conceptual analysis. In the words of Bealer:

> [A]lthough [the scientific evidence] bears on "intuition" in an indiscriminate use of the term, they evidently tell us little about the notion of intuition […] which is relevant to justificatory practices in logic, mathematics, philosophy and linguistics.[34]

Now, it doesn`t follow that the *rational* intuition is infallible. We just don`t have any reason to believe that it is - in normal circumstances – *systematically* misleading.[35] The critics of conceptual analysis might concede the point here and claim that even the non-systematic unreliability suffices to disqualify *rational* intuition. Are they right? There is a convincing argument against this weaker claim of the critics. The argument was formulated by Ernest Sosa. Sosa draws an analogy between perception and *rational*[36] intuition. He reminds us of what we all agree on: both perception and *rational* intuition are fallible, unreliable and corrigible. However, neither of them are *systematically* so. The analogy doesn`t finish here. Perception is the very starting point of any *empirical* investigation. Similarly, *rational* intuition is the very starting point of any *theoretical* investigation. We have an obvious inconsistency here: although perception and *rational* intuition seem to suffer from the same unreliability it is only the *rational* intuition that is – by some - rejected as worthless. What Sosa says about perception applies without qualification to rational intuition:

[33] For a detailed discussion of the *rational* intuition and how it differs from other varieties of intuition (that is, those that are the focus of scientific research) see George Bealer, `Intuition and the Autonomy of Philosophy`, in Michael R. DePaul and William Ramsey (eds), *Rethinking Intuition* (Oxford: Rowman & Littlefield Publishers, Inc., 1998), pp.201-239.

[34] Bealer, *Rethinking Intuition*, p.213.

[35] It could be argued that it is impossible in principle to demonstrate plausibly and convincingly that *rational* intuition is – in normal circumstances – systematically misleading. The basic idea behind such an argument would be something like this: any attempt (theoretical or experimental) to demonstrate the systematic unreliability of rational intuition could always be traced back to some (implicit) initial rational intuition that constitutes the starting point of that very attempt to demonstrate its systematic unreliability. And that`s clearly self-defeating.

[36] Sosa doesn`t use the expression `rational intuition`, just `intuition`. However, it is clear, from the context, that when he talks about intuition, he means something sufficiently close to Bealer`s `rational intuition`.

> It is evident that human perception is fallible. More, it is known that human perception is systematically misleading in certain conditions. What should we conclude about our faculty of perception? In conditions known to psychologists, a normal subject would be systematically misled unless aided by collateral information. Before discovering ways in which perceptual conditions can be misleading, one is liable to go astray systematically in a variety of perceptual beliefs. Absent collateral information, human perception falls far short of epistemic perfection. […] In the light of this, what do we conclude about the epistemic value of perception? Surely it would be precipitous and imperceptive to condemn perception wholesale on the basis of such fallibility. Would it not be comparably precipitous and imperceptive to condemn intuition wholesale?[37]

Let me sum up the defence of intuition. The scientific evidence referred to by the critics doesn`t focus on *rational* intuition. Thus, the criticism of intuition based on that evidence misses the target. An advocate of conceptual analysis admits that *rational* intuition is fallible. She argues, however, that (a) the fallibility of *rational* intuition is not relevantly different from that of perception and (b) similarly to perception, the *rational* intuition is irreplaceable. To be consistent – the advocate will insist – we either reject *both* perception and *rational* intuition as epistemically worthless or none of the two. And, of course, the point is that no one is ready to ditch the former.

Now, I agree with the objection and concede that, strictly speaking, there are no necessary and/or sufficient conditions and no conceptual truths available to an a priori investigator. Agreeing with the critics here, however, does no harm to the particular conceptual analysis I engage in above. The reader will have noticed that when discussing the conceptual connection and the nature of the dependency between *free will/agency* and *moral responsibility*, I am not searching for any necessary and/or sufficient conditions to associate with the concept of free will/agency. I engage in something much less ambitious. Instead of seeking to identify the necessary and/or sufficient conditions, I am seeking to identify a conceptual *confusion*. The conceptual confusion I am after above is the implicit or explicit assumption that free will/agency has essentially something to do with moral responsibility.[38] Understanding and practising conceptual analysis as seeking to identify conceptual *confusions* is – unlike searching for necessary and/or sufficient conditions - an entirely plausible project. Consider the following analogy: it might be futile to try to identify the necessary and/or sufficient conditions for the concept-applicability of *consciousness* and yet one will be very safe to rule out that it has anything relevant to do with the concept of, for instance, liquidity. To claim the contrary, would surely strike one as a case of serious conceptual confusion. However, not all conceptual confusions are similarly transparent. Often, they are hiding behind superficial

[37] Ernest Sosa, `Minimal Intuition`, in in Michael R. DePaul and William Ramsey (eds), *Rethinking Intuition* (Oxford: Rowman & Littlefield Publishers, Inc., 1998), pp.257-269 (p.261).

[38] Notice that this critical focus of mine on the assumed or stipulated essential connection between *free will/agency* and *moral responsibility* is perfectly in line with the claim of my potential critics that there are no essential and/or sufficient conditions.

plausibility and their identification will require a conceptual analysis of the kind I have advocated and performed above.[39]

1.3 The second step

It should be clear at this point that the concept of free will is independent from the concept of moral responsibility. As such, any theorizing about it should proceed outside the conceptual constraints associated with *moral responsibility*. Theorizing within the constraints is seriously misguided and will lead to a theoretical impasse. Below, I will show to what impasse the theorizing within the constraints associated with *moral responsibility* can lead.

1.3.1 Two constraints on the concept of free will

There are, in contemporary free will literature, two major approaches to the concept of free will. The first one understands free will as being primarily a function of the availability of alternative possibilities, while the other one takes free will to be primarily a function of an agent being the ultimate source of her actions. Robert Kane introduces both approaches in the following passage:

> We believe we have free will when we view ourselves as agents capable of influencing the world in various ways. Open alternatives, or alternative possibilities, seem to lie before us. We reason and deliberate among them and choose. We feel (1) it is "up to us" what we choose and how we act; and this means we could have chosen or acted otherwise. As Aristotle noted: when acting is "up to us," so is not acting. This "up-to-us-ness" also suggests (2) the ultimate control of our actions lie in us and not outside us in factors beyond our control.[40]

Traditionally, the overwhelming majority of philosophers working on the problem of free will have taken the availability of alternative possibilities to be the core condition of the sort of freedom required for justified attributions of moral responsibility. In words of John Martin Fisher:

> [T]he most influential view about the sort of freedom necessary and sufficient for moral responsibility posits that this sort of freedom involves the availability of genuinely open alternative possibilities at certain key points in one's life.[41]

A growing minority of philosophers focus on the second approach arguing that the most fundamental feature of a morally responsible agent is her ability to – in some appropriate way - \`originate\` or be \`the source\` of her actions. They will insist that the attributions of moral

[39] In the next chapter, when formulating an argument, I will make use of the notions of necessary and sufficient conditions. I will do so being, at the same time, fully aware of the limitations described above. The conclusion and implications of the argument will be presented as modally strong but nowhere close to a logical necessity.

[40] Robert Kane, *A Contemporary Introduction to Free Will* (New York: Oxford University Press, 2005), p.6.

[41] John Martin Fisher, \`Recent work on moral responsibility\`, *Ethics*, 110, no.1 (1999), pp.93-139 (p.99).

responsibility cannot be justified if `the sources of an agent`s actions do not originate *in* the agent but are traceable to factors outside her`.[42]

Now, within the framework of theorizing about freedom-as-moral responsibility,[43] it seems to be plausibly possible to drop the requirement dictating that an agent must have genuine alternatives available to her if she is to be treated as morally responsible. The existence of this plausible possibility has been demonstrated by Harry Frankfurt in his seminal paper `Alternate Possibilities and Moral Responsibility`.[44] In the paper,[45] Frankfurt formulates a thought experiment that, slightly modified,[46] reads as follows:

> Suppose Black wants Jones to perform a certain action. Jones is inclined to perform the action but hasn`t yet fully committed to performing it. Black prefers Jones to decide on his own to perform the action. However, Black is ready to interfere should Jones decide not to perform the action. Black secretly inserts a chip in Jones`s brain. Now Black can monitor and control Jones`s behaviour. Black`s intention is this. He will monitor Jones`s brain. If Jones decides to perform the action, Black will not interfere. However, should Jones show any inclination to refrain from performing the action, Black will interfere and manipulate Jones into performing it.

The thought experiment invites the reader to test their conceptual intuition. If Jones decides on his own to perform the action, is he morally responsible for it? The majority of philosophers agree that, at least on the face of it, Jones should be held responsible for performing the action if it was his own decision. At the same time, it looks like Jones doesn`t really have an option to do otherwise. Even the slightest intention to refrain from performing the action will be detected by Black, who will then interfere and make Jones perform the action anyway. Thus, the thought experiment seems to show that one can be morally responsible without having alternatives available to them. Some philosophers have been convinced by the thought experiment, others not.[47] What matters for my purposes here is that these Frankfurt-type thought experiments cast serious doubts on whether freedom-as-moral responsibility requires the agent to have alternatives available to her. It, then, seems safer (and it is the only option

[42] Michael McKenna, `Robustness, control, and the demand for morally significant alternatives: Frankfurt examples with oodles of alternatives`, in David Widerker and M. McKenna (eds), *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (VT: Ashgate, 2003), pp.201-217 (p.201f).

[43] Below, I shall use the expression `freedom-as-moral responsibility` to emphasize the reference to the kind of freedom that will support attributions of moral responsibility.

[44] Harry G. Frankfurt, `Alternate Possibilities and Moral Responsibility`, *The Journal of Philosophy*, 66, no.23 (1969), pp.829–839.

[45] Ibid., p.835.

[46] The original version of the thought experiment leaves it unspecified how exactly Black is going to ensure that Jones ultimately does what Black wants him to. I take the advantage here of the state of contemporary neuroscientific research that makes it perfectly plausible to talk about brain implants as means of controlling one`s behaviour. I don`t think the same level of plausibility was available to Frankfurt when he published the article in 1969.

[47] A good overview of arguments for and against the conclusion Frankfurt draws from the thought experiment is provided by J. M. Fisher in his `Frankfurt-Type Examples and Semicompatibilism: New Work`, in Robert Kane (ed.), *The Oxford Handbook of Free Will*, 2nd edn (Oxford: Oxford University Press, 2011), pp.243-265.

left to those who accept Frankfurt`s interpretation of the thought experiment) to try to build a successful account of freedom-as-moral responsibility around the concept of sourcehood. There is no thought experiment showing – in a way analogous to the Frankfurt-type scenario – that it can be (or can feel) right to attribute moral responsibility onto an agent that is *not* a source of her actions. It won`t surprise then that the overwhelming majority of philosophers see sourcehood as a necessary condition of freedom-as-moral responsibility. These philosophers will understand the problem of free will/agency as having the form of the following question: how can someone be a source – in the sense of being a point where the causal origins of something truly start - of an event in a causally determined world?

Below, I wish to cast doubt on the very prospects of solving the problem of free will/agency so understood. Before I proceed to do so, let me remind the reader what the background dialectics is here. There is nothing *obvious* about *sourcehood* being something like a necessary condition of free agency. It comes to the foreground only once *freedom* gets qualified as *freedom-as-moral responsibility*. When searching to identify the one *responsible* for an action, we are *intuitively* searching for the *source*: the one with whom the (causal) buck stops. Thus, whatever theoretical difficulties the logic of the concept of sourcehood involves comes in our way solely due to the stipulated qualification.

1.3.2 The troubles with *sourcehood*

So, what are those theoretical difficulties the logic of the concept of sourcehood involves? And how serious are they? What follows is a brief outline of the theoretical trajectory (and where it ends) that any critical inspection of the concept of sourcehood in the context of a morally responsible agency will likely follow. First, the uncontroversial stuff. It is widely assumed that any successful account of free agency will relate actions to their agent in such a way that it will be correct to describe the agent as the `source` of those actions. The assumption is not of a *philosophical* nature but simply spells out what any competent speaker understands as entailed in the expression `agent`s actions`. A competent speaker will understand the phrase `these are actions of a particular agent` as synonymous with the phrase `a particular agent is the source of these actions`.

Let`s stay in a non-philosophical territory for a bit longer. A dictionary will give us something like the following definition of `source`: *the place something comes from or starts at, or the cause of something.* Now slightly more formally: `source` is an expression that identifies the referent of a concept or a proposition *x* as being the spatiotemporal or causal origins of the referent of a concept or a proposition *y*. Examples:

Beer is a good source of vitamin B.

> *We walked up the creek to its source in the mountains.*
>
> *Extreme inequality has been identified as the main source of widespread obesity in the developed countries.*

The first example informs us that one of the origins of, or the *places* to look for, vitamin B is beer. The next one gives us the *location* the creek starts at and the last one identifies an unexpected originating *cause* of a particular health issue. Now, there is something fairly uncontroversial about sourcehood to notice here. In each of the examples the referent of \`sourcehood\` could be replaced by a description referring to objects or events that are (causally) constitutive of the replaced referent. The referent of this replacing description could then plausibly be seen as the true referent of \`sourcehood\`. Thus, the \`beer\` (as an expression referring to a source of vitamin B) could be replaced by a description of relevant biochemical processes that result in vitamin B formation in beer. Similarly, the expression referring to the particular location in the mountains where the water of the creek is said to have its source can be replaced by a description that will identify as the true source all the various ways via which the water gets to that particular location, e.g. subterranean stream(s), surface water and precipitation. Finally, \`extreme inequality\` (as an expression referring to the source of widespread obesity) could be replaced by a description capturing the explanatory causes of extreme inequality such as, unchecked financial deregulation, tax havens and a lack of efficient redistributive mechanisms. In all three cases, the replacing description refers to a new sourcehood.

Notice that the replacing description in our examples is not meant to preserve synonymity with the replaced one. The replacing description plays an explanatory role. It plays the explanatory role in virtue of reinterpreting the original referent in terms of prior (causal and spatiotemporal) events or phenomena that constitute it. The *prior* constitutive events referred to in the replacing description play the role of a new sourcehood. This identifying of the *prior* constitutive events can be seen as moving the sourcehood backwards on the spatiotemporal-plus-causal axis. Now, there are practically no limits to this backward motion. Any sourcehood can be pushed backwards on the axis and replaced by events and/or phenomena constitutive of that sourcehood.[48] Of course, we don\`t normally question the utility of the notion of a source on the grounds that there isn\`t really any source because it can always be pushed back on the axis. We will, typically, accept something as a source of something else if it provides the required richness of explanation for that something. And that required richness will vary depending on the context. A casual reader might accept that the source of the tragic death of

[48] Two exceptions come to mind: God and the Big Bang. None of the two is open to reinterpretation in terms of prior constitutive events and/or causes. The reason is that they both, by definition, kick-start the time flow so there cannot be any prior events associated with them.

72 people in Grenfell Tower block of flats was the fire started by a malfunctioning fridge-freezer on the fourth floor, while a social scientist might see as the true source of the tragedy the decision of the local council to use cheaper but flammable cladding.

Three claims should feel uncontroversial at this point:

C1: A source of something can always be pushed back on the spatiotemporal-plus-causal axis.

C2: A different concept or description will be used to refer to this new – pushed back – source.

C3: For practical purposes, we fix the source (of something) on the spatiotemporal-plus-causal axis at a point which gives us the required explanatory richness.

How does this all relate to sourcehood as discussed in the context of moral responsibility discourse? As discussed above, the logic of justifiable attributions of moral responsibility is subject to an intuitive requirement that the agent is the true source of her actions, i.e. that she is where the (causal) buck stops when traced back. In other words, there is an intuitive requirement that an agent is morally responsible only if the source of her actions cannot be traced beyond her agency. I shall call this requirement a *non-traceability condition* (NTC). Now, (C1) tells us that the sourcehood can always be pushed back on the spatiotemporal-plus-causal axis. The agency of a person and the sourcehood ascribable to it exist on the axis. This strongly suggests that (C1) applies here and the sourcehood can always be pushed and traced beyond the agency. At the same time, (C3) tells us that, normally, our choosing among the candidates for the source of something is guided by explanatory utility each of the candidates comes with. The practicalities of the everyday social interaction with others are best approached under the assumption that me and the others are the source of their actions. Thus, we have strong *practical* reasons to understand our own and the others` agency as the locus of sourcehood of actions. However, the problem of free will/agency is, ultimately, a *metaphysical* problem. As such, it dictates to find out how things *really* are and not how it is *useful* to see them; that is, it dictates to disregard (C3). This leaves us with (C1) and its immediate implication, which is that: the sourcehood of an action can always be pushed and traced beyond the related agency.

Now, there is no way to block the implication once (C1) is accepted. This leaves a potential challenger with a desperate move: to reject (C1). The move gets us deep into the territory of the unintelligible. The heart of the claim in (C1) is that anything that exists has a causal history. The events constitutive of a particular causal history themselves have a causal history. And *that* causal history is, of course, again constituted by prior events with their own causal history and so on and so forth. Denying this amounts to claiming that there can be causally efficacious events that do *not* have any causal history. Accepting that such events can exist gives us a theoretical framework within which the concept of agency can meet the constraint of (NTC). Our agency will be declared capable of originating actions in a way that is not (fully) causally embedded in the state of the world prior to those actions. Agents will be thus understood as capable of an act of originating or causing that itself has *not* been originated or caused.

The idea of `originating that itself is not originated` is not new. Aristotle, Aquinas and others would recognize the phrase as referring to the concept of *causa sui*: a cause that causes itself.[49] It might not be immediately obvious but the notion of `a cause that causes itself` is an extreme one. On the face of it, it seems to consist of concepts that we have a good grasp of: *a cause*, *causation* (towards), *itself*. Thus, we might, reading the phrase, get a feeling that we have a grasp of something here. We haven`t though. Even a rather casual reflection exposes an unyielding unintelligibility of the notion. How does one as much as begin to conceive of something that causes itself? A philosopher as deep as Nietzsche was so appalled by the notion of *causa sui* that he called it `a sort of rape and perversion of logic`.[50] Much more recently, Galen Strawson has argued convincingly that moral responsibility is ultimately impossible because it would require the agent to be *causa sui*, a requirement that Strawson finds so obviously absurd that without much supporting argument he rests his case confident that something like an ultimate moral responsibility is metaphysically impossible.[51] It won`t surprise then that this truly mindboggling capacity of self-causation has been attributed to God only.[52] I shall take it as uncontroversial that *causa sui* is an impossible concept and reaffirm (C1).

1.4 Conclusion

Let me, briefly, sum up what has been done above in steps 1 and 2 and how the steps relate to each other. In step 1, I drew the reader`s attention to the fact that in the literature on free will the concepts of free will and moral responsibility are – implicitly or explicitly – understood as standing in a relationship of *mutual* conceptual dependence. I argued that, at best, the

[49] The concept of *causa sui* entails the absence of causal history.

[50] Friedrich Nietzsche, *Beyond Good and Evil*, Walter Kaufmann (trans.) (New York: Vintage, 1966), § 21.

[51] Galen Strawson, `The Impossibility of Ultimate Moral Responsibility` in Galen Strawson, *Real Materialism and Other Essays* (Oxford: Clarendon Press, 2008), pp.319-336.

[52] Consider for instance Aristoteles` concept of *prime mover* – the first uncaused cause of all the motion in the universe; a being that is perfectly beautiful, indivisible, intelligent and self-contemplating. For more see Aristotle, *The Metaphysics*, Hugh Lawson-Tancred (trans.) (London: Penguin Classics, 2004), book Lambda. Similarly, process theologians such as Charles Hartshorne, Jon B. Cobb, Jr. and David Ray Griffin who – inspired by process philosophy of Alfred Whitehead – will see God as *causa sui*. For more see John W. Cooper, *Panentheism: The Other God of the Philosophers* (Grand Rapids: Baker Academic, 2006).

dependency runs in one direction only: the concept of moral responsibility theoretically depends on the concept of free will but not vice versa. I proceeded to argue against a possible suggestion that it is a permissible move to simply *stipulate* that *free will* conceptually depends on moral responsibility. I claim that such a stipulation is seriously misleading and that, consequently, reaching a theoretical impasse should be expected. Step 2 is meant to show how the conclusion of step 1 applies to the current state of discussion in the literature on free will. I briefly presented the reasons behind my strong suspicion that any theorizing about free will that continues to take place within the constraints of moral theory will, ultimately, lead to an unintelligible requirement on any account of free will/agency that aspires to succeed. As a result of an unwarranted stipulation of conceptual dependence of *free will/agency* on *moral responsibility*, any successful account of free will/agency will have to be built around the concept of an agent that is capable of uncaused causation. The notion of uncaused causation is a seriously incomprehensible one. Any theoretical account built around such a notion is, I am convinced, doomed to fail.

Steps 1 and 2 relate to each other in the following way: the argumentative burden behind my call to sever the concept and discussion of free will/agency from the concept and discussion of moral responsibility is carried by step 1. Step 2 provides motivating considerations by showing the seriousness of the theoretical impasse that the arbitrary stipulation of conceptual dependence of *free will/agency* on *moral responsibility* leads to.

Chapter 2: Freedom and reasoning

2.1 Introduction

In this chapter, I will argue for the following claim:

C: Freedom is attributable only to agents who exercise practical reasoning.

An outline of the argument:

i. Free will/freedom is attributable to agents only.
ii. *Freedom in agency* entails the agent having a sense of an alternative open to her.
iii. *Having a sense of an alternative* - where the notion of an alternative is such that it sustains attributions of *freedom* - entails *having a sense of being able to choose*.
iv. *Choosing* entails *deliberation*.
v. *Deliberation* entails *an exercise in reasoning*.
vi. *Freedom in agency* entails *an exercise in reasoning.*

2.2 Free will/freedom is attributable to agents only

In Chapter 1, I argued that tying the discussion of free will with that of moral responsibility is confused and misleading. We need a different starting point. I suggest starting from (i) as such a different starting point. (i) states that: free will/freedom[53] is attributable to agents[54]

[53] We have a bit of the paradox of analysis here. I am trying, ultimately, to arrive at a novel understanding of freedom (in agency). To avoid circularity this novel understanding cannot be presupposed in (i). At the same time freedom in (i) must have some positive content if we want (i) to have a truth value. In other words, the problem is that the content cannot involve any crucial elements of the novel understanding of freedom (to avoid even partial circularity), neither can it contradict the novel understanding if we want to avoid equivocation in the usage of *freedom* throughout one and the same argument. A way to neutralize the paradox is, roughly, this. There are two kinds of understanding: a propositional one and a non-propositional one. Both can be about the same subject matter. A person can have one without the other. You can non-propositionally understand how to generate a forehand loop when playing tennis to keep the ball in the court without having any propositional understanding of how the imparted spin causes the ball to curve downwards. And vice versa, you can perfectly understand the physics behind the downward curve of a spinning ball without being able to generate one, i.e. without having a non-propositional understanding how to hit a spinning ball. Applied to (i): As competent users of the concepts of free will and freedom we have an intuitive and competent understanding of the concepts. The understanding is, at least partially, non-propositional. I don`t intend to come up with any novel modifications to this non-propositional understanding. My target is the propositional understanding, and this is where I intend to be original. Importantly, there is no circularity or paradox here as `understanding` is being used in two different senses. Those wishing to learn more about the paradox of analysis, see Colin McGinn`s excellent treatment of it in the fourth chapter of his *Truth by Analysis: Games, Names, and Philosophy* (Oxford: OUP, 2011), pp.47-61.

[54] The concept of agency is a rather heterogenous one. The Oxford English Dictionary mentions as the earliest use of `agent` to be found in a treatise on alchemy, written in 1471, where it refers to `a force capable of acting on matter`. This usage seems to have survived until the present time in connection with, for instance, chemical weapons. The toxic component of a chemical weapon is called

only. The claim is just a first approximation and its function is purely negative: to block attributions of freedom/free will onto such clearly freedom-lacking entities as stones, trees, ponds, clouds, etc. The claim introduces a very general necessary condition of attributability of freedom. The reader could concede that, indeed, we wouldn`t want to attribute free will/freedom to the above-mentioned entities while, at the same time, denying that it has anything to do with agency. She could simply suggest that the reason we wouldn`t want to attribute freedom/free will to those entities has to do with them showing no symptoms of anything like non-determined behaviour or a behaviour that would involve self-causation (*causa sui*). Such a suggestion would be wrong though. Consider the behaviour of a subatomic particle. A subatomic particle is a non-agential[55] entity that exhibits an intriguing behaviour. Physics tells us that any future state of a subatomic particle is undetermined; i.e. it tells us that a subatomic particle is a non-deterministic entity in the sense of having alternative future states open to it. The non-deterministic nature of a subatomic particle also makes it interpretable as a *causa sui* phenomenon. The fact that a subatomic particle has alternative future states open to it is intelligible only if it entails that nothing outside of it determines (causally or otherwise) its future state. The remaining option is that the particle determines its future state from within itself.

Now, the availability of alternative possibilities and/or the possession of the *causa sui* power should, at least on the face of it, feel to us as sufficient to justify attributions of freedom/free will. Indeed, it is the very understanding of *freedom/free will* in terms of these conditions what constitutes the problem of free will in the first place, because it is these conditions that seem to be deeply irreconcilable with the logic of causal determinism.[56] [57] And yet, despite the fact that a subatomic particle seems to be describable in terms constitutive of the concept of freedom/free will, we would find it strongly counterintuitive

its `chemical agent`. This is, however, not a kind of `agent` we would wish to attribute free will/freedom onto.

[55] If subatomic particles were agents, then virtually anything in the physical world would have to count as an agent in virtue of being an aggregate of agents (unless one would want to deny the existence of collective agency). The distinction between agents and non-agents would collapse.

[56] There are influential compatibilist strategies arguing that the availability of alternative futures is not a necessary condition for attributability of freedom (Harry Frankfurt`s and his follower`s). The central claim of these compatibilist accounts of freedom doesn`t really stand in any tension with what I am saying above. These compatibilist accounts of freedom are the result of a philosophical argument motivated by an effort to solve the problem of free will. This effort involves initially accepting the dilemma of the problem of free will; i.e. it involves accepting freedom as having essentially something to do with alternative futures and/or *causa sui* power.

[57] The concept of *causa sui* is not, strictly speaking, in tension with the theory of causal determinism. It is in tension with a broadly related physicalist principle called The Principle of the Closure of the Physical. This, however, doesn`t affect the logic of the problem of free will at all. The ultimate puzzle – how can free will exist in the physical reality as we know it? – remains to be solved. For more see Robert C. Bishop and Harald Atmanspacher, `The Causal Closure of Physics and Free Will`, in Robert Kane (ed.), *The Oxford Handbook of Free Will* (Oxford: Oxford University Press, 2011), pp.152-170.

to characterize a subatomic particle as `free` or as possessing `free will`.[58] Clearly, there must be something else, another condition, that needs to be met to justify an attribution of freedom/free will. And it is against the background of these considerations that I suggest – as a new starting point - that the attributability of freedom has essentially something to do with agency.

The claim that the attributability of freedom has essentially something to do with agency is just a first approximation. Certain constitutive aspects of agency will have to be in place for *freedom* to be attributable. Consider, for instance, a sleeping agent who is having a dream. When such a sleeping and dreaming agent moves her hand in her sleep, we wouldn`t want to say that moving her hand was an exercise in agential *freedom*. We wouldn`t want to say this even after it was pointed to us that the sleeping agent`s hand movement was accompanied by certain quite complex mental processes, i.e. a certain richness of consciousness and emotional responses that dreaming comes with. This simple consideration suggests that, (a), an unqualified agency will not do when it comes to attributability of freedom and, (b), that not just *any* mental states will sustain the attributability.

So, what exactly is it about agency that needs to be in place for the attribution of freedom to hold? I am arguing towards the claim that freedom is attributable to agents only if they, in some suitable way[59], act rationally. I could, perhaps, ground the claim in the appeal to some of the insights generated by the contemporary discussions (see footnotes 56 and 57 below) regarding the structure of actions. These discussions seem to agree, at least implicitly, on certain basic features an action must have to count as one. Now, what can be said about the nature of an action applies to the nature of agency. The reasoning here is that: an agent constitutes itself *by* acting or ceases to be an agent entirely; *action* and *agent* are the two sides of the same conceptual coin. Thus, the structural features identified as necessary for something to be an action will be the same as those necessary for agency to be one. And if freedom can be ascribed *only* to agents – which is the claim in (i) - then the necessary structural features of agency (and actions) will have to be understood as the necessary conditions for the attributability of freedom itself. So, what can we learn from the contemporary philosophy of action?

[58] It could be objected that while it is rather counterintuitive to ascribe *free will* to a subatomic particle it is much less so if what we ascribe to it is *freedom*. It is perhaps true that it is less counterintuitive to attribute *freedom* to a subatomic particle then it is to attribute *free will* to it. Still, we wouldn`t want to do it as it comes at a price. We possess concepts of *chaos* and *randomness*. The two concepts feel very suitable for describing the behaviour of a subatomic particle. If we allow *freedom* to refer to the same aspects of that behaviour, we end up with equating *freedom* with *chaos* and *randomness*. And that can`t be right.

[59] I will say more about what counts as a *suitable* way of acting rationally below. To quite some degree, however, I will leave it up to the reader to bring into the picture their own understanding of rationality. My account of freedom in agency can accommodate widely differing theories of rationality.

The central focus in the philosophy of action is on the necessary conditions of applicability of the concept of action, i.e. it is about answering the question, `What does it take to be an action at all?` or, alternatively, `What distinguishes action from mere activity or mere behaviour?`. The answers given share an appeal to an underlying logical form that an action must conform to if it is not to lapse into something less than that, i.e. into a mere activity or behaviour. What is meant by `underlying logical form` here? A good way of getting the idea here is to have a brief look at two popular views of action: the calculative and the authorship views.

The most prominent advocates of the calculative view of action are Candace Vogler and Michael Thompson. According to Vogler,[60] the internal structure of actions is a series of steps towards a termination point. The individual steps in the series relate to each other in a rational way, i.e. a step is rational if it is a step on the way towards the termination point of the action that you are in the course of performing. The steps themselves are actions organized by coherent rationality striving to reach the final stopping point. The performance of the individual steps is motivated by *calculative* reasons, that is, reasons whose force is: this is a step toward the termination point of my action. For Vogler, *calculative* reasoning is what holds the internal structure of actions together. If missing or deficient then the action collapses into a mere activity or behaviour. Thompson is saying something very similar: actions have other, `smaller`, actions as their parts; an action is always constituted by modules that are themselves actions.[61] These modules consist of further submodules that, again, are themselves actions. Our grasp of inference is what organizes nesting of actions in a way that leads to a successful performance of a target action. The laws of inference constitute the underlying structure.

The authorship view of action doesn`t seem, on the face of it, to be about identifying an underlying structure as an essential feature of an action. Instead, it starts from a plausible assumption that any action must have an owner (similarly, any belief must have one). Thus, what matters, according to proponents of the view, is that an action is *authored*. One of the most developed versions of the authorship view has been formulated by Christine M. Korsgaard. On Korsgaard`s view, the essential feature that distinguishes action from mere activity is that actions – unlike activities – are owned.[62] An action can be owned only if the owner is a whole person, i.e. only when the action is produced by the agent as a whole. The behaviour that is not produced by the agent as a whole but by her psychic parts is nothing more than events that just resemble actions. To be a whole person requires a *constitution*, a

[60] See Candace Vogler, *Reasonably Vicious* (Cambridge: Harvard University Press, 2002).
[61] See Michael Thompson, *Life and Action* (Cambridge: Harvard University Press, 2008), part II.
[62] For Korsgaard`s views see Christine M. Korsgaard, *The Constitution of Agency* (Oxford: OUP, 2008).

certain form of psychic organization. For the required constitution to obtain, the agent – when producing the action – must identify[63] with a principle of choice, where that principle must be universal in form. Consequently, the reasons guiding the choice, too, must be universal in form.[64] If the agent fails to adhere to the universal forms of practical reasoning her authorship is compromised and `her` action collapses into a mere event.

The brief sketch of the theories of action given above plays a very modest role in my argument. It is briefly mentioned here just to show that it is not uncommon among philosophers working on the theory of action to think about action and agency as having something essentially to do with an exercise of reasoning. It would be perhaps acceptable to fully ground (C) – that is, the central claim of this chapter - in the appeal to those theories only. There are, however, two reasons why I prefer not to do so. First, the theories mentioned here seem to share a serious weakness. They all are – at the core – stipulative and/or prescriptive about the necessary conditions of action. That is, rather than telling us what actions (and agents) *are*, they tell us what they *should* be. The central claim of this chapter is one of the crucial steps in my argument. Relying on a stipulative and/or prescriptive element would make the argument *as a whole* similarly stipulative and/or prescriptive. Thus, grounding (C) in those theories only would render my whole argument unacceptably weak. Second, I believe that (C) can be grounded in a robust conceptual argument that will keep (C) and the rest of the argument on a metaphysical path. Steps (ii) – (vi) constitute such an argument.

Before I proceed, let me briefly address an important conceptual issue here. Step (ii) makes a frequent usage of the notion of entailment. The notion of entailment implies necessity. The implied necessity in the notion of a conceptual entailment commits one to the talk of a conceptual *necessity* and/or conceptua*l truth*. Typically, a conceptual investigation yields, among its results, a proposition *x entails y*. The proposition is meant to express a conceptual truth or a conceptual necessity regarding (the concept) *x.* Here the conceptual truth is that whatever *x`s* extension is, *y* must, in some sense, be a part of it. There are, however, powerful and, as it seems to me, conclusive arguments against the viability of the notions of *conceptual necessity* and *conceptual truth*. In his inaugural address,[65] Timothy Williamson argues convincingly that there is nothing necessary about our concepts or their mutual relations. This applies, among others, to the notion of conceptual entailment which,

[63] In this aspect, Korsgaard`s account resembles that of Harry Frankfurt. Frankfurt argues that a desire (and the related action) is full-fledgedly yours when via your second-order desire (the desire that targets the first-order one) you identify with the first-order one.

[64] At this point Korsgaard takes a broadly Kantian path and insists that the reason deployed in the choice of action must satisfy certain (Kantian) universal-in-form principles.

[65] See Timothy Williamson, `Conceptual Truth`, *Aristotelian Society Supplementary Volume*, 80, no. 1 (2006), pp.1-41.

of course, represents one such relation. I won`t go into detail about Williamson`s argument here. I am perfectly happy to accept Williamson`s conclusion as nothing in my argument turns on it or its refutation. At the same time, I shall continue using the notion of entailment throughout the thesis, and do so for two reasons. First, there is no ready to hand alternative to the notion of entailment that does the job without implying necessity. Second, throughout my conceptual investigation, I will be after conceptual *implications* and *connections* whose modal tie to truth is, though not necessary, sufficiently strong. Most of our philosophical theorizing is perfectly satisfactory even if grounded merely in strong intuitive plausibility and/or strong overall coherence. To search for *necessary* conceptual relations is not only – if one agrees with Williamson – a futile project but also one that philosophers don`t have to pursue. Thus, I shall ask the reader to remember that the notion of entailment (and the related notions of conceptual necessity and conceptual truth) as used throughout the thesis doesn`t imply necessity[66] but only something like a strong modal tie to truth.[67]

2.3 *Freedom in agency* entails the agent having a sense of an alternative open to her

I believe it uncontroversial that in our intuitive, common sense, understanding of *freedom* we take it to entail the availability of an alternative.[68] I don`t intend to establish – via an empirical survey perhaps – whether this belief of mine is true or not. The reason for this lack of intention is that even if this claim about *freedom* turned out to be untrue we – who work on the problem of free will - would have to assume otherwise. We would have to assume otherwise as the very entailment is what co-constitutes the original dilemma at the heart of the problem of free will. Recall the logical structure of the dilemma: The world is causally determined. Causal determinism excludes an availability of alternatives. We are *free* agents. *Freedom* entails an availability of alternatives. Thus, *free* agency entails a condition that is excluded by causal determinism. Clearly, if *freedom* did not entail this condition (i.e. the

[66] There still might be adherents of *conceptual necessity* and/or *conceptual truth* out there*;* and judging by the not infrequent occurrence of those notions in contemporary philosophical literature there are quite a few of them. These adherents can read *entailment* as involving necessity should they wish to do so. The logic of the argument stays unaffected; just its modal force is greater.

[67] At the core of this is our faculty of *rational* and *modal* intuition. This is a crucial and foundational faculty in any philosophical and other theorizing about the world. My views here on the *rational* and *modal* intuitions and their epistemological role are more or less indistinguishable from those of George Bealer (see, for instance, his `Intuition and the Autonomy of Philosophy`, in Michael R. DePaul and William Ramsey (eds), *Rethinking Intuition* (Oxford: Rowman & Littlefield Publishers, Inc., 1998), pp.201-239.). I won`t go into more detail and defend the view here as nothing of importance turns on it. My argument will retain sufficient force even if the notion of (conceptual) entailment as used throughout the argument is understood as packing just enough modal force to convince most of the competent users of the relevant concepts. Perhaps this force won`t be particularly strong, just stronger than that of the alternatives.

[68] That is, having a sense of *at least* one alternative is entailed here. Of course, it could be more. The point here is that one can be seen as a free agent only if one is having a sense of at least being able to *refrain* from a particular course of action (where *refraining* from a particular course of action constitutes an *alternative*).

condition of an availability of alternatives), we wouldn`t have any problem of free will to theorize about in the first place.

It could be objected that this dilemma is not the whole story. The problem of free will is often recognized as involving two dilemmas. The one mentioned in the previous paragraph turns crucially on understanding *freedom* as entailing availability of alternatives. But does the other too? The other dilemma is, roughly, this: We can be said to exercise our free agency only if we have *control* over the available alternatives. The world is causally *undetermined*. *Agential control* over available alternatives entails a capacity to *determine* the choice of the preferred alternative. In the world that is *essentially* undetermined, it is a mystery how such a capacity of *determination* in choosing could be embedded in it. Here the dilemma is rising out of the conceptual entailments of the concepts of *control* and *determination*: controlling is determining and determining is controlling. Thus, it might seem that the problem of free will survives in a version that is independent of understanding *freedom* as entailing availability of alternatives. If true it would contradict my claim that the logical structure of the problem of free will commits us to understanding *freedom* as entailing an availability of alternatives. It will, however, be immediately noticed that the concept of *agential control* is about control over *available alternatives*. It is so simply because the concept of control entails (the availability of) alternatives as its target. There must be a possibility for things to go a different way if left uncontrolled for the notion of control to make any sense. The *possibility for things to go a different way* is synonymous with *availability of alternatives*.

There is a complication related to the notion of control that has to be addressed here. In their *Responsibility and control: A theory of moral responsibility*, Fischer and Ravizza exploit the logic of a Frankfurt-type scenario to distinguish two kinds of control: regulative control and guidance control. The regulative control involves the dual power to perform either an action X or some other action Y. That is, the regulative control involves availability of alternative possibilities. The guidance control, on the other hand, doesn`t involve any such power. Fischer and Ravizza illustrate the distinction on the following example:

> Sally is driving her car with a driving instructor sitting next to her. The road they are on is winding up a steep hill. Sally is carefully guiding the car through the many twists and turns the road has. The driving instructor is quite happy with Sally`s performance and lets her guide the car on her own. However, should Sally show any inclination to steer the car off the road, the driving instructor will interfere, take control of the car and keep it on the winding road.[69]

The example of Sally and the driving instructor shows that there is a kind of control that doesn`t involve the availability of alternative possibilities. Undeniably, it is Sally who

[69] This scenario is a short and paraphrased version of the case presented in John Martin Fischer and Mark Ravizza, *Responsibility and control: A theory of moral responsibility* (New York: Cambridge University Press, 1998), pp.30-32.

controls the car when *guiding* it through the twists and turns of the road. At the same time, Sally doesn`t have the option to deviate from the bounds of the road curvature. Any deviation will immediately be stopped by the driving instructor. Sally controls (has *guidance* control of) the car in the sense of *making* it follow a predetermined trajectory having, however, no (*regulative*) control over *what* trajectory her car will follow.

As mentioned above, the regulative control involves the availability of alternatives while the guidance control doesn`t. Thus, Sally can be plausibly described as having control of her car without having any (relevant) alternatives available to her when controlling it. Clearly, this contradicts the above made claim of mine that the concept of control entails (the availability of) alternatives. How serious is this? Not very, I believe. Let us concede, for the sake of the argument, that Fischer and Ravizza have demonstrated successfully that something like guidance control – that is, a kind of control that doesn`t involve availability of alternatives – is a conceptual possibility.[70] However, and importantly, Fischer and Ravizza draw the distinction between the regulative and the guidance control in the context of theorizing about moral responsibility (and not in the context of theorizing about free will). Fischer and Ravizza are determinists. They deny the existence of free will. This denial, they claim, doesn`t imply the non-existence of moral responsibility because moral responsibility can be grounded in guidance control, which is a capacity of agents that is fully compatible with determinism. The reader will remember that in the first chapter we have emancipated theorizing about free will/agency from theorizing about moral responsibility. This means that we should be wary of the conceptual intuitions elicited within the context of moral theory. Outside of that context, the concept of free will/freedom feels to be undeniably about availability of alternative possibilities. Thus, although the notion of control might, perhaps, not involve alternatives when analysed within the context of theorizing about moral responsibility, it certainly does so when analysed within the context of theorizing about free will/freedom. (By the way, the reason Fischer and Ravizza deny the possibility of free will - but not the possibility of moral responsibility - is exactly because they are determinists who take it as obvious that the notion of free will implies alternative possibilities.) We can conclude, I believe, that the very recognition of freedom (in agency) or free will as a

[70] It is, in fact, a hotly contested issue whether the Frankfurt-kind cases (like the one with Sally and the driving instructor) show that the notion of control without alternatives is a coherent one. Perhaps the most serious objection against the concept of guidance control comes from (rather surprisingly) John Martin Fischer himself. Fischer argues that the very possibility that an agent might try an alternative constitutes what he calls a `flicker of freedom` (see his *The Metaphysics of Free Will: An Essay on Control* (Oxford: Basil Blackwell, 1994), p.134-147). Kadri Vihvelin argues, as I understand her, along a similar line (see her `Freedom, Foreknowledge, and the Principle of Alternate Possibilities`, *Canadian Journal of Philosophy,* 30, no. 1 (2000), pp.1-23 (pp.8-9). For an intriguing argument against the `flicker of freedom` strategy, see Eleanor Stump, `Alternative Possibilities and Moral Responsibility: The Flicker of Freedom`, *The Journal of Ethics*, 3, no. 4, The Contributions of Harry G. Frankfurt to Moral Responsibility Theory (1999), pp.299-324.

philosophical *problem* commits one to understanding[71] *freedom* as entailing the availability of alternatives.

Now, there are some distinctions I wish to make about the notion of *availability of alternatives*. Making the distinctions will expose some ambiguities in the notion allowing me to arrive at the claim in (ii). Alternatives can be available in three ways:

W1. Metaphysically only. The agent is not aware of them.

W2. Both metaphysically and in awareness.

W3. In awareness[72] only. Metaphysically unavailable.

(W1) is, on the face of it, a perfectly intelligible kind of availability (of alternatives). It refers to a rather common state of affairs: something is the case out there in the world without the subject being aware of it. Thus, it feels perfectly intelligible to say, in many situations, that there were options, possibilities or alternatives without the agent (or anyone else) being aware of them. Above, when discussing subatomic particles and alternative future states available to them, I assumed an agreement on subatomic particles being incapable of any awareness. I expected no controversy when describing them as having alternatives (alternative future states) available/open to them. Let`s take (W1), for now at least, to be a clear and uncontroversial distinction even though below I will argue that (W1) is problematic and, ultimately, incompatible with *freedom*.

Regarding (W2), I am unable to think of a problem or an objection here. If it is agreed that (W1) is clear and uncontroversial then (W2) must be even more so as there is a sense in which the state of affairs referred to in (W2) is a more basic state than the one referred to in (W1). Clearly, to be able to conceive of a scenario where something is metaphysically available without the relevant subject being aware of this availability, one must be first capable of conceiving a more primitive scenario: a scenario in which one *is* aware of metaphysically available alternatives. The point here is that there is no way one could arrive at thinking about the world as containing available alternatives that one is unaware of without prior acquaintance – that is, without prior experience in awareness – of some of the available ones. How could we think about the world as containing any alternatives had we never been

[71] This is not to say that this understanding of *freedom* has to be preserved in the final account of *freedom*. It might turn out – as a result of theoretical analysis – that *freedom*, to be a coherent notion, must be divorced from the assumption(s) that was/were thought of as constitutive of it at the beginning of the theoretical inquiry.

[72] Below, when discussing (W3), I will switch to talking about an agent *having a sense* of alternatives instead of an agent being *aware* of alternatives. Being *in awareness* implies the existence of that which is in awareness (as *to be aware* is a factive verb), which is an implication I wish to avoid in (W3). I present the distinction in (W3) in terms of (being in) *awareness* because that`s the notion used in (W1) and (W2) and its usage in (W3) makes it easier to see how the three ways that the alternatives can be given relate to each other.

aware of any? Thus (W2) is a default position here while (W1) depends – in a way that doesn`t need to be explicated here – on it.

To see how best to understand (W3), let`s first have a look at how (W1) relates to (W2). (W2) is, as explained above, in some sense, the most basic way that alternatives can be given: There is an alternative available to me to act on and I am aware of it. I have run out of milk, so I go to a local grocery shop to get a box. When about to pay, I reach for my debit card that I always carry in the left front pocket of my jeans. It is not there and I immediately realize I left it on the desk in my study room. I get annoyed as I will have to go back home for it and then repeat the trip to the grocery shop. I don`t stay annoyed for long though. Still at the counter I recall, suddenly and with relief, that there is some change in one of the pockets of my coat - just enough to buy the milk. I don`t have to go back home for the debit card. I have an alternative: I can pay by cash that is in my coat pocket, and I am aware of this alternative. This is a case of (W2).

Cases like this and their analogues are very common. Perhaps even more common is the following variation. The story runs an identical course up to the point when I realize that I have left my debit card at home. I don`t happen to recall that there is some cash (enough to buy the milk) in one of my coat pockets. I return home for the debit card. Only there I remember the money in the coat pocket. I get rather annoyed. I could have paid using the cash in my coat pocket and could have saved myself a boring walk back to the shop. The alternative was there but I was not aware of it. This variation is a case of (W1).

Now, (W3) too is derived from the more basic (W2). Consider yet another variation of the milk story. I need to go to a local grocery shop to get a box of milk. I can be rather absent minded so, still at home, I check my pockets. The debit card is in my jeans and there is some change (three pounds and a few coppers) in my coat. I suspect that there might be insufficient funds on my debit card but I don`t worry about it much as there is an alternative: I could pay by cash. Unbeknown to me the coat pocket where my cash is meant to be has got a little hole in it. Putting the coat on I stretch it a bit and the hole in the pocket gets bigger; big enough for the coins to slip through. I lose the coins when walking on a grassy path in the park that is situated between my house and the grocery store. I didn`t hear the coins hit the ground as the ground is soft and grassy. At the till in the grocery shop, it turns out there are insufficient funds on my card and – to my puzzlement – there is no change in my coat pocket. It seems natural to describe the scenario as one in which, from a certain point on, the alternative of paying by cash existed in my awareness *only*.

Now, clearly, we could think of many other scenarios analogous to the milk story and its variations. The milk story kind of scenarios are a rather common part of our everyday life. This commonality is what gives (W1) – (W3) a strong intuitive support. There is, however, a problem regarding (W1) and an issue to be clarified regarding (W3).

Recall the claim in (W1): alternatives can be available metaphysically *only*, that is, they exist out there in the world without the agent being aware of them. In what sense, however, can we talk about *freedom* here if the kind of availability of alternatives it entails is of the (W1) kind? Consider the following scenario:

You are locked up in prison and you would like to escape. One of the bricks in the wall of your cell is loose. It can be taken out. There is a little cavity in the wall behind the brick with keys from the prison in it. You have not got a slightest idea about the loose brick, the cavity, the keys. There is an alternative to your remaining locked up in the prison. You could take the keys, open the doors and walk out.

Yet, it seems extremely unintuitive to describe you as being *free* on the grounds that there is a relevant alternative to your staying locked up in the cell available to you in this way, i.e. in the (W1) way.

Now, it is immediately obvious that the reluctance to describe you in this scenario as being *free* (with respect to your stay in prison) has something to do with the lack of *awareness* of the relevant alternative. Our theoretical response to this can be one of the following: we could either deny that something can be an *alternative* without being, at the same time, in awareness, or insist that the notion of *availability* implies being in *awareness*. For our purposes here it doesn`t matter much which response one prefers as both of them boil down to an identical necessary condition to be associated with the *availability of alternatives* in the context of the entailment of this notion in the concept of freedom. The identical necessary condition is this: attributing freedom to an agent requires the agent to be *aware* of at least one[73] of the relevant[74] available alternatives. Thus, it turns out, under closer inspection, that (W1) refers to a kind of availability of alternatives which is incompatible with *freedom*.

[73] There is also a complication here regarding possible redundancy of relevant available alternatives. There could be two cavities with identical keys in the cell with the agent knowing about one of them only. Or there could be, instead of keys and unbeknown to the agent, enough money to bribe the guards. Thus, it could be objected that the agent is not aware of (some other or all) relevant available alternatives and the necessary condition for ascribing freedom is thus not met. At the same time, it feels intuitively correct to attribute *freedom* to him just on the grounds of him knowing about the cavity with keys. A simple qualification will suffice here: The agent must be aware of *at least* one relevant and available alternative.

[74] Clearly, there might be unrelated alternatives that the agent is aware of. Awareness of unrelated alternatives is irrelevant. Consider the prison-cell-loose-brick-keys-in-the-cavity scenario again. The prisoner desires to escape the prison. Imagine there is another loose brick with a cavity in the wall of the cell. Instead of a set of keys there is a flute in the cavity. The prisoner happens to know about this. It could be suggested, as an alternative to using the keys he knows are in the other cavity, that he takes out the flute and learns to play it. In the context of his desire to escape the prison this alternative is irrelevant. His unawareness of this irrelevant alternative does not affect the attribution of *freedom* onto him once it is conceded that he knows about the cavity with the keys.

We are left with (W2) and (W3) as the two remaining ways of how, in the context of attributions of freedom, alternatives can be available. Our discussion of (W1) above has resulted in identifying a condition that must be met for the notion of availability of alternatives to sustain attributions of *freedom.* We know now that an agent must be *aware* of a (at least one) relevant available alternative, i.e. it will not suffice for the alternatives to be available metaphysically *only*. The condition of an agent`s *awareness* of available alternatives turns out to be a necessary one.

Now, this talk of *awareness* of available alternatives seems to imply that available alternatives are, somehow, out there and we are or are not aware of them. The notion of *awareness* refers to a *factive*[75] mental state. Awareness is always *of* something that *really* is the case. If that something doesn`t *really* exist neither does the awareness of that something. Thus, any talk of *awareness* of alternatives implies that the alternatives really are out there in the world. If this is correct then (W3) is an incoherent claim because it posits in awareness metaphysically unavailable items, which is at odds with the *facticity* of *awareness*. There is a relatively easy way of dealing with this issue about (W3). We can switch from talking about *awareness* of alternatives to talking about *sense* of alternatives. The semantics of *having a sense of* something doesn`t require for that something to really exist out there as *having a sense of* something refers to a *non-factive* mental state. Thus, an alternative could be said to be available in a (W3) way if an agent had a *sense* of it even though there wouldn`t, as a matter of fact, really be an alternative out there in the world. This little terminological move gets rid of the contradiction in (W3).

Now, there is an important thing to notice here. It will often be *phenomenologically indistinguishable* to an agent whether the relevant alternative is given in (W2) or (W3) way; i.e., it will often be impossible for the agent to know or be sure whether it is just a *sense* of an alternative she is having or *awareness* of it.[76] Recall the variation of the milk story where I first make sure – become aware - that there is cash in my coat pocket only to discover later

[75] A good way to understand what factive mental states are is to contrast them with the non-factive ones. In words of Jennifer Nagel: `Factive mental states, such as knowing or being aware, can only link an agent to the truth; by contrast, non-factive states, such as believing or thinking, can link an agent to either truths or falsehoods.` See Jennifer Nagel, `Factive and non-factive mental state attribution`, *Mind and Language,* 32, no. 5 (2017), pp.525-544 (p.1).

[76] I offer the following consideration in support of the viability of the notion of phenomenological indistinguishability as related to the distinction between a *sense* and *awareness* (of an alternative): Were we able to determine whether what we are having is a *sense* or *awareness* of an alternative, we wouldn`t find the problem of free will terribly exciting. We would simply be able to say, i.e. to *distinguish*, whether the world *is* or *isn`t* such that it contains real alternatives. We would be able to do so just by inspecting our mental states. That is, if we, in one case at least, established that what we are having is not a *sense* but *awareness* of an alternative, we would be able to conclude that we are, in principle, *free* agents. I am convinced that it is exactly this, in principal, inability to determine whether what we experience is a *sense* or *awareness* of an alternative that is a necessary precondition of us recognizing the problem of free will as a problem at all.

that, to my surprise, there is none (as there is a hole in my pocket and the coins have slipped through it unnoticed). It would feel extremely implausible to suggest that I was somehow able to notice – perhaps by being attentive to my internal mental state - the switch from having *awareness* of the alternative to having a *sense* of it (after the coins slipped through the hole in my pocket).

There is an important asymmetry here. *Being aware of an alternative* entails *having a sense of an alternative* but not vice versa. The reason is that *having a sense of an alternative* refers to the phenomenological component of the mental state of having *awareness* of an alternative.[77] Thus, whenever one is *aware* of something, he is, at the same time and inseparably, having a *sense* of that something, i.e. having a *sense* of an alternative is a necessary component of having *awareness* of it. *Having a sense of* something, on the other hand, doesn`t entail the existence of that something and as such doesn`t imply *awareness.* What does it mean for our discussion of the concept of freedom/free will as entailing availability of alternatives? It means that having a sense of an alternative is, due to it being an essential component of any relevant availability of alternatives, a necessary condition for attributability of freedom/free will. In other words: *freedom in agency* entails the agent having a sense of an alternative open to her.[78]

2.4 *Having a sense of an alternative* – where the notion of an alternative is such that it sustains attributions of *freedom* - entails *having a sense of being able to choose*.

We have the same propositional phrase – *having a sense of* – on both sides of the entailment claim. *Having a sense* entails *having a sense.* This is a trivial and uninformative kind of entailment. The interesting entailment is the one that, as I claim, obtains between *an alternative* and *an ability to choose*. This entailment is not an obvious one. Recall the story that physics tells us about subatomic particles. According to the story, a subatomic particle has, at any point, a number of alternative future states open to it. Each state open to it comes with a certain probability value attached to it. The value is higher than zero.[79] At the same time, a subatomic particle is not capable of *choosing* anything. What we have here is a plausible talk of genuine availability of alternatives that is accompanied by the lack of an

[77] The other component of the mental state of *awareness* of an alternative (or anything else) is the alternative itself (or anything else) out there in the world.

[78] A reminder: at this point it seems natural to take (W2) as the only kind - out of the three - of availability of alternatives that will sustain attributions of *freedom* onto agents. In this context, my conclusion that *freedom in agency* entails the agent having a sense of an alternative open to her must be understood as identifying that which cannot be given up, i.e. that which is *necessary* for attributability of *freedom.* Nothing that has been said *so far* indicates that having *awareness* of alternatives (i.e. the availability of *real* alternatives) is an unnecessary condition and can be given up. Only later, and as a result of a dedicated argument, it will be shown how *freedom* can be plausibly attributed to agents even without metaphysical availability of alternatives.

[79] And, of course, lower than one for the other states to have some probability of obtaining.

ability to choose. Thus, it looks like we can talk perfectly well about alternatives without implying an ability to choose (one of them). This doesn`t look good for my claim that *having a sense of an alternative* entails *having a sense of being able to choose*. A closer inspection is needed.

Let me derive a definition of *an alternative* from the subatomic particle story:

DA: Something is an alternative if and only if the probability of it obtaining is higher than 0 (and lower than 1).

Now, what we are after in our conceptual analyses of *freedom in agency* is a notion of an *alternative* that will sustain attributions of freedom. Thus, we have to ask: Can *an alternative* as defined in (DA) do the job of sustaining such attributions? It can`t. Consider the following scenario:

The world is such that there is a set of states of affairs involving me that could happen two minutes from now where each state has a positive (and at the same time lower than 1) probability of obtaining. A brilliant scientist (helped by a supercomputer) comes up with a list of those states. He gives it to me. I read it and become aware of all the alternatives that are open to me in this way. However, I am incapable of choosing – and aware of it - any of those states and have to wait for the two minutes to pass to learn which of the alternatives – that involve me – will obtain.

I expect no objections whatsoever to refusing to describe myself as *free* in that scenario. Such a description is simply way too counter-intuitive despite the fact that there is a robust sense in which I have alternatives open to me in that scenario. The scenario exposes two things. First, it shows that the concept of an alternative as defined in (DA) won`t sustain attributions of *freedom*. Second, it shows that the notion of an alternative that is needed to sustain the attributions must entail *an ability to choose.*

Within the phenomenological framework of *having a sense*, the above will work in the following way. When, in a particular situation, I wish to establish – via reflection – whether or not I am *free*, I will identify as relevant only those alternatives I have a sense of that come merged with me having a sense of being able to *choose* them. How do I know that this is how the identification will proceed? It has been tested above. In the scenario above, I – and the reader – have already tested our conceptual intuition regarding the entailment relation running from *freedom* to *alternative* to *choosing*. When reflecting on whether or not I am *free* in a particular situation, I will employ – on pain of being inconsistent – the same conceptual intuition and identify something that I have a sense of as an *alternative* only if it comes merged with me having a sense of being able to *choose* that something.

I can conclude now: *having a sense of an alternative* – where the notion of an alternative is such that it sustains attributions of *freedom* - entails *having a sense of being able to choose*.

2.5 *Choosing* entails *deliberation*

What is *choosing*? I am, ultimately, looking for an account of freedom in agency built around a notion of freedom that can be recognized by a non-philosopher. The concepts constitutive of such a notion should themselves be similarly recognizable. In this context, the safest way is to have a look at what an authority such as the Cambridge English Dictionary has to say about *choosing*. The dictionary[80] gives the following two definitions of the verb *to choose*[81]:

- According to British English, (BE), *to choose* is: *to decide what you want from two or more things or possibilities*
- According to American English, (AE), *to choose* is: *to think about which one of the several things is the one you want, and take the action to get it*

It is immediately obvious that more is involved in the meaning of the verb as understood in American English compared to its understanding in British English. In American English, the verb *to choose* entails *taking an action* while there is no such entailment essential to it in British English. At the same time, both (BE) and (AE) take the verb to refer to a process of picking out – from two or more options – what one wants. British English specifies the process as *deciding* what one wants while American English specifies it as *thinking about which one* it is that one wants. The phrases *deciding what one wants* and *thinking about which one it is that one wants* refer to an identical process.[82] Now we can express the relation between (AE) *choosing* and (BE) *choosing* in something like the following simple way: (AE) *choosing* = (BE) *choosing* + a relevant action (or, alternatively: (BE) *choosing* = (AE) *choosing* – a relevant action).

I wish to convince you that *choosing* entails *deliberation*. This task requires focusing on what is conceptually essential[83] about *choosing* and disregarding that which is inessential.

[80] See <https://dictionary.cambridge.org/dictionary/english/choose?q=to+choose> [accessed 16 September, 2019].

[81] I shall be using *choosing* and *to choose* interchangeably depending on the syntactical requirements of the sentence it is being used within.

[82] Why do the two phrases refer to an identical process? *Choosing* is defined in terms of these two phrases. To preserve the definitional equivalence, the verb *to choose* and the two phrases must have the same meaning. But they can`t have the same meaning if they have different referents.

[83] The notion of the *essential* is to be understood here in a relatively weak sense. The standard understanding is that if *x* is *essential* to *y* then *y* has *x necessarily*. Above (p.25), I conceded – partly because doing so had no bearing on my argument – that there are no conceptual necessities. If true, then there can`t be any conceptual essences either. Thus, analogously to my treatment of the notion of *entailment* (above, p.25), I shall disconnect the notion of *essential* from that of *necessity* and associate it with the condition of *strong intuitive plausibility* in something like the following way: *x* is *conceptually essential* to *y* if – in the absence of a successful counterargument – the majority of

The little `semantic` equation above identifies the *action* as an inessential element. This element cannot be conceptually essential to *choosing* as it is not implied by the (BE) understanding of it. The essential bit here is the shared reference to the process of picking out an option. And the conceptualization of this process is what we need to have a look at to be able to see that and how *choosing* entails *deliberation*.

The process that *choosing* refers to is conceptualized either as *deciding what one wants* or as *thinking about which one it is that one wants*. Put simply, *choosing* has essentially something to do with *deciding* and/or *thinking.* Below I will want to argue for quite an uncontroversial claim that both *deciding what one wants* and *thinking about which one it is that one wants* are phrases whose meaning overlap substantially with that of *deliberation*. Prior to that, however, there are issues that need to be addressed.

2.5.1 *The exclusivity requirement*

There is a way of understanding *deciding* (and *thinking*) which could be seen as a problem for my argument. To discuss this, an assumption needs to be introduced. The assumption is plausible and its function is to put a constraint on any successful account of freedom (in agency).

> *The exclusivity requirement*: An account of freedom (in agency) that allows attributions of freedom onto entities lacking phenomenological states is to be deemed unsuccessful.

I assume agreement on the claim that we would not want robots and other purely algorithmic systems to be understood as *free* in any but most metaphorical sense. *The exclusivity requirement* is meant to keep it that way. Two things need to be briefly mentioned here. First, the assumption has, I believe, quite a strong *prima facie* plausibility. For whatever reason – be it the fact that robots follow preprogramed algorithms, or that they lack feelings, or that they are not biological entities - it just feels strongly unintuitive to attribute freedom to a robot or other purely algorithmic system. Our conceptual intuitions here regarding the attributability of *freedom* should be taken seriously exactly because the problem of *freedom* is a *conceptual* problem. Second, I will say things, below, that go beyond a mere appeal to intuitive plausibility regarding the assumption. When explaining how exactly my argument is kept within the bounds of *the exclusivity requirement* I will, at the same time, provide *reasons* for accepting it.[84]

competent speakers with a good grasp of *x* and *y* find it *strongly intuitively plausible* that the semantic extension of *y* includes that of *x*. This notion of the *essential* is relatively weak compared to the one that implies necessity. It is, however, strong enough to sustain a convincing conceptual analysis.

[84] There might be readers who don`t find attributions of *freedom* onto robots counterintuitive and who, at the same time, won`t find my argument in support of the *exclusivity requirement* convincing. Those readers will then read – and correctly so – my account of freedom (in agency) as permitting attributions

Now, as mentioned above, there is a problematic understanding of *deciding*. What makes this understanding problematic is that it could, if not neutralized, set my account of freedom onto the path leading up to the violation of *the exclusivity requirement*. In other words, there is an understanding of *deciding* (and of *thinking*) that would ultimately turn my account into an account that allows attributions of freedom onto robots and other purely algorithmic systems. Such an account of freedom would be implausible and unsuccessful.

To see what the troubling issue about *deciding* is, consider the following. Suppose that there is a robot equipped with a system of artificial intelligence which, among others, is built so as to satisfy *modus ponens*. Thus, whenever the robot gets into a state that counts as endorsing a representation of its environment such that it can be captured by propositions of the form \`*p*\` and \`if *p*, then *q*\`, it will move to a state that will count as endorsement of a representation of its environment that has a propositional form \`*q*\`. Now, is it correct to say that the robot *draws inferences* according to the law of *modus ponens*? There is a sense in which it does and a sense in which it doesn\`t. Clearly, the robot is able to form and endorse a new representation of its environment which conforms to the inference rule of *modus ponens*. In this sense, the robot is capable of drawing inferences. On the other hand, and equally clearly, the robot doesn\`t draw inferences in that *rich* way that logicians do because it lacks *awareness* of the rule that governs it. The *awareness* of the rule is what allows a logician to *apply* the rule, while the lack of it is what makes the robot to *conform* to it.[85] *Deciding* is similarly ambiguous. A robot, or some other purely algorithmic system, processes an input in accordance with a set of rules and *decides* what output it will generate. It doesn\`t feel too unnatural to describe the robot as *deciding* even though the process of deciding would be understood as purely algorithmic, i.e. as a process lacking any awareness or a phenomenological component. At the same time, it is obvious that *deciding* as conforming to purely algorithmic rules is importantly different from *deciding* as a process that essentially involves a phenomenological state (of awareness).

Now, the problem is roughly this. If *choosing* can be construed in terms of *deciding* and *deciding* can be understood as a purely algorithmic process, then *choosing* too can be understood as a purely algorithmic process. *Choosing*, on the face of it and very intuitively, is a core notion in any plausible account of freedom (in agency). Freedom entails availability of alternatives, and being, among other things, *choosable* is what makes something an alternative. Thus, if *choosing* can refer to a purely algorithmic process, it opens the door for

of freedom onto robots. That shouldn\`t discourage them from accepting my account because attributions of freedom onto robots is something that they don\`t see as counterintuitive and problematic anyway.

[85] The distinction between *applying* a rule and *conforming to* a rule is due to Phillip Pettit. See his \`Deliberation and Decision\`, in Timothy O\`Connor and Constantine Sandis (eds), *The Companion to the Philosophy of Action* (Wiley-Blackwell, 2011), pp.252-258 (p.254).

justified attributability of freedom onto robots. That would violate *the exclusivity requirement* and render my account of freedom unsuccessful.

One might suggest sticking with the sense of *deciding* that essentially involves a phenomenological state (of awareness). That would restrict the applicability of *choosing* to systems with phenomenological states thus keeping the argument within the bounds of *the exclusivity requirement*. Unfortunately, we don`t have the liberty to choose the sense of *deciding* that best suits our purposes because the logic of the argument requires that we respect all relevant conceptual intuitions. Throughout my argument I rely heavily on the relation of conceptual entailment[86], peeling off the individual layers of conceptual entailment and trying, in this way, to arrive at the target claim that *freedom in agency* entails *an exercise in reasoning*. The entailment in the target claim is meant to be sufficiently strong – that is, grounded in a sufficiently strong conceptual intuition - which can be the case only if the steps of the argument preserve the strength. Thus, once it has turned out that *deciding* could be taken as a purely arithmetic process – that is, once it has turned out that *deciding* is not essentially a phenomenologically loaded process – we are bound to accept it into our conceptual analysis.

It could, perhaps, be suggested that we will have more luck with keeping the argument within the bounds of *the exclusivity requirement* if we drop talking about *choosing* in terms of *deciding* and replace it with talking about *choosing* in terms of *thinking.* The reason for the replacement would then be that thinking might seem to be – when compared to deciding – harder to conceptualize as a process completely devoid of any phenomenology. And if thinking is essentially a phenomenological process then choosing is too. Defining *choosing* in terms of a phenomenologically loaded notion of *thinking* would thus keep the argument within the bounds of *the exclusivity requirement*. This won`t work for two reasons. First, there is, among contemporary philosophers, no consensus about the putative phenomenological nature of thinking. Philosophers of naturalistic inclination impressed with the theoretical models of cognitive neuroscience will deny that thinking essentially involves a phenomenological component. Naturalists are, typically, reductionists about phenomenology whenever at all possible. A naturalist will find it perfectly plausible to understand *thinking*, along with *deciding*, as notions referring to a purely computational process. And naturalism is too popular to dismiss out of hand its view that there is nothing essentially phenomenological about thinking (or indeed about anything else). Second, and more importantly, even those who – like cognitive phenomenologists[87] – will find it clear

[86] In the relatively weak sense of *entailment* as explained above (see pp.24-25).

[87] Perhaps the most prominent advocate of cognitive phenomenology is Galen Strawson. See for instance his `Cognitive Phenomenology: Real Life` in Tim Bayne and Michelle Montague (eds), *Cognitive Phenomenology* (Oxford: Oxford University Press, 2011), pp.286-325. The philosophers who have kick started the cognitive phenomenology project are Terence Horgan and John L. Tienson.

and obvious that thinking is phenomenologically loaded, won`t be able to take this path out of the problem. They won`t be able to do so for exactly the same reasons that have been presented above in connection with the notion of deciding; i.e. specifically in connection with the two readings – an arithmetic and a phenomenological - that the notion permits. We couldn`t, for reasons given above, opt for the phenomenological reading and discard the arithmetic one. Analogously, we can`t choose *thinking* over *deciding* in our analysis of *choosing* just because it suits the argument. If *choosing* can be defined in terms of *deciding* then it must be accepted as essential to it and accepted along with all the theoretical consequences.

It doesn`t look good. Plausibly, both *deciding* and *thinking* can be understood as referring to a purely arithmetic process. Even if it is somehow successfully argued that *thinking* is essentially a phenomenological notion, it won`t help. The logic of the argument requires that the relevant conceptual entailments run in one direction without any forking that would support a different claim. The very *possibility* (and plausibility) of defining *choosing* in terms of (a purely arithmetic notion of) *deciding* pushes the argument onto a collision path with *the exclusivity requirement.* What can be done about this?

It will be noticed that the discussion above of *choosing, deciding* and *thinking* has been done from a certain perspective. Let me distinguish two perspectives here:

> *A pn-perspective*: a phenomenologically neutral perspective, and
>
> *A pl-perspective*: a phenomenologically loaded perspective

I draw here on a familiar distinction between the third-person and first-person views of mental events in the philosophy of mind. *A pn-perspective* would then be a perspective that shares its vantage point with that of the third-person view. And *a pl-perspective* would then be a perspective that shares its vantage point with that of the first-person view. Clearly, the perspective from which our discussion above of *choosing*, *deciding* and *thinking* has been done is the *pn-perspective*. Now, how does the distinction and its application help us here? Recall (iii):

> *Having a sense of an alternative* – where the notion of an alternative is such that it sustains attributions of *freedom* – entails *having a sense of being able to choose*.

See their seminal `The Intentionality of Phenomenology and the Phenomenology of Intentionality`, in David Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary readings* (Oxford: Oxford University Press, 2002), pp.520-533. For an excellent collection of essays on cognitive phenomenology see Tim Bayne and Michelle Montague (eds), *Cognitive Phenomenology* (Oxford: Oxford University Press, 2011).

Here the talk of *choosing* is embedded in the *phenomenological* context of *having a sense*; i.e. here *choosing* is to be looked at from the *pl-perspective*. The subsequent discussion defined *choosing* in terms of *deciding* and *thinking*. Thus, if *choosing* is to be looked at from the *pl-perspective* so are *deciding* and *thinking*.

Now, looking at the two notions and the processes they refer to from the pl-perspective will give us an interesting conceptual constraint regarding the understanding of the two notions. Above we have seen that it doesn`t feel terribly unnatural to talk about *deciding* and *thinking* in connection with purely algorithmic systems. That stops being the case once we switch the perspectives. Reflect for a moment on what`s happening when you are having a *sense* of thinking. Are you having a sense of an algorithmic process? You might be tempted to give an affirmative answer. There is, however, a strong reason to believe that such an answer would not be correct. The notion of an algorithmic process is a notion that refers to a *theoretical* (not empirical) object. As a theoretical object it can be experienced (or given in one`s sense) only as a *thought*. It cannot be allowed that the thought itself, in some higher order awareness, is given as an algorithmic process. An appeal to *thinking* would have to be made on that higher order level too, and that would start an infinite regress. Thus, in one`s sense – or from the *pl-perspective* – thinking is always given as phenomenologically loaded, or not given at all. Exactly the same goes for deciding. If deciding is assumed to be a purely algorithmic process than it follows that it is available in one`s sense as a *theoretical* object. Theoretical objects can be experienced as *thoughts* only. To avoid an infinite regress those thoughts cannot – in the domain of one`s sense - be reduced to an algorithmic process only. Looking at deciding from the *pl-perspective* reveals that deciding too involves a phenomenological state.

Let`s sum it up. The possibility of *deciding* and *thinking* being taken as referring to a purely arithmetic process set my argument on a collision course with *the exclusivity requirement*. A distinction between a phenomenologically loaded perspective - *a pl-perspective* - and a phenomenologically neutral one - *a pn-perspective* - was made. The analysis of *deciding* and *thinking* that allowed for the notions to be conceived of as referring to a purely arithmetic process was identified as conducted form the *pn-perspective*. However, my argument discusses *choosing* and the related notions of *deciding* and *thinking* as embedded in the phrase - *having a sense of*, i.e. it discusses the notions from the *pl-perspective*. It turns out that from the *pl-perspective* the notions of *deciding* and *thinking* (and *choosing*) cannot be conceived of as referring to a purely arithmetic process. That will do for the argument to meet *the exclusivity requirement*.

2.5.2 *Deciding* and animal agency

I wish to convince you that *choosing* entails *deliberation*. Although I believe that the claim is quite uncontroversial it adds force to my overall argument if I break the claim into conceptually even more transparent and plausible steps. Part of this has been done above. I have defined and discussed *choosing* in terms of *deciding what one wants* and *thinking about which one it is that one wants*, which has moved us into a close proximity of *deliberation*. *Deliberation* is, in one of its senses, *weighing reasons* or *considering* and *comparing* the features of the available options (on the basis of which the agent is led to a choice). Now, *weighing reasons* or *considering* and *comparing* the features of the available options is just nothing but *thinking which one it is* in the context of determining the preferable option. That is, *deliberation* is *thinking which one it is* that is the preferable one and vice versa: *thinking which one it is that one wants* is *deliberation* applied in the context of determining the preferable option. That much seems to be clear. What is, however, less transparent is the relation between *deliberation* understood as *weighing reasons* and/or *considering* and *comparing* the features of the available options and the notion of *deciding what one wants.* It might not be immediately obvious how, or whether at all, *deciding* entails *weighing reasons* and/or *considering* and *comparing* the features of the available options.

The issue here is closely related to the one regarding animal agency. We often employ intentional verbs when explaining behaviour of higher animals (i.e. those that display certain minimal levels of behavioural complexity). Such an employment of intentional vocabulary will feel most natural in the case of apes and cetaceans due to their high level of intelligence. However, we will often feel unconstrained to do the same even in the case of relatively lower animals such as dogs or cats. Consider the answer to the following question: `Why is your dog barking at me?` Your reply could be something like: `Because she *thinks/believes* that you want to steal food from her feeder.` That would be a perfectly natural answer. And if attributing *thinking* and *believing* to dogs and cats feels perfectly natural, the same must be the case with attributions of *deciding* as there seems to be no reason to assume that deciding requires a more complex neural base for its instantiation than thinking and/or believing does. True, the talk of dogs *deciding* anything might perhaps feel less natural then the talk of them *thinking* or *believing*, but it still is far from unacceptable. Consider the title of an article in The Huffington Post: `Dog Given A Home To Die In ... But He Decides To Live Instead …`[88] Surely, that sounds alright. Now the problem here is that although we might be ok with dogs *deciding* things, we also might want to resist any talk of dogs *weighing reasons* and/or *considering* and *comparing* the features of the available options.[89] This would suggest that

[88] Caitlin Jill Anders, 'Dog Given a Home to Die In ... But He Decides to Live Instead', (Huffpost, 2016).<https://www.huffingtonpost.co.uk/entry/dog-near-death-lives-in-new-home_us_56951b9ee4b05b3245da6a3a> [accessed 23 February 2018].

[89] It could be objected here that we waived the option of resisting the talk of dogs *weighing reasons* and/or *considering* and *comparing* features of available options when we accepted as perfectly natural

deciding doesn`t, after all, entail *weighing reasons* and/or *considering* things. If this is correct, I might not be able to conclude that *choosing* entails *deliberation* because: *deliberation* is *weighing reasons* or *considering* things, and if *choosing* entails *deciding* (as established above) without *deciding* entailing *weighing reasons* and *considering* things – that is, without *deciding* entailing *deliberation* – then I have failed to make the crucial step[90] towards the claim that *choosing* entails *deliberation*.

So how serious is it? Not very, I believe. The force of the objection stems from the *prima facie* plausibility of taking dogs as deciding things without taking them, at the same time, as being capable of weighing reasons and/or considering things. That would imply that *deciding* has nothing essentially to do with *weighing reasons* and/or *considering* things. There is an inconsistency and a confusion at the heart of the objection though. To address the inconsistency and the confusion, I need to mention briefly certain issues discussed by philosophers concerned with animal agency.

Some of the central questions the philosophers theorizing about animal agency ask are: Can non-human animals act intentionally? Can they act for a reason? Unsurprisingly, some philosophers participating in the discussion of these questions will give an affirmative answer while others will give a negative one. Both sides will, however, agree on something like this: intentional behaviour is a behaviour that is both motivated by *reasons* and that involves an ability to, in some sense, *operate* with *reasons*. Now, the answer given to those two questions above will depend on what one takes to count as acting for a *reason* or *operating* with *reasons*. More specifically, it will depend on how demanding one`s conditions for the attributability to a creature of (acting for) a reason or operating with reasons are. You can, for instance, endorse *lingualism* tying the attributability to the possession of language. Lingualists – Davidson and others[91] – will deny on a priori grounds

to attribute *thinking* to dogs in sentences such as: `The dog was barking at me as she *thought* that I was about to steal food from her feeder`. Surely – the objector would insist – weighing reasons and/or considering and comparing features is an exercise of thinking. Thus, we should find it similarly natural to talk about dogs weighing reasons and/or considering and comparing features. True, weighing reasons and/or considering and comparing features is an exercise of thinking. It doesn`t follow, however, that the two mental processes (thinking on one side and weighing reasons and/or considering and comparing features on the other) are identical and that the relevant notions can be used interchangeably. The notion of thinking can refer to quite a basic mental process/state compared to the one referred to by *weighing reasons* or *considering*. Thus, when we say `He *thinks* Berlin is the capital of Germany`, we don`t imply – regarding the complexity of the mental process referred to by the employed intentional verb - more than when saying that `He *believes* Berlin is the capital of Germany`. Clearly, thinking in the sense of believing is much more basic than weighing reasons or considering things. And that explains why there is no conceptual dissonance in attributing to dogs thinking while, at the same time, resisting to see them as weighing reasons or considering features.

[90] A step in the transition of entailment.

[91] For instance: P.M.S. Hacker, *Human Nature: The Categorical Framework* (Oxford: Blackwell, 2007), pp.204, 236, 240; John McDowell, *Mind and World* (Cambridge, MA: Harvard University Press, 1996), p.70; Harry Frankfurt, *The Reasons of Love* (Princeton: Princeton University Press, 2004), pp.18-19.

that animals without language can have mental capacities at all, or at least the `higher` mental capacities that are required for rational agency. This is a very demanding condition because, arguably, it is only humans that possess language. Consequently, only humans are capable of acting for a reason and operating with reasons. Another related strategy would be to tie rationality with conceptual competence and argue that animals lack conceptual competence on the grounds that they fail to meet a requirement known as the *generality constraint*. The generality constraint was introduced by Gareth Evans[92] as a constraint on genuine concept possession, and thus as a constraint on a creature's capacity for authentic thought. Someone capable of a thought that *a* is *F* has two distinct capacities. She must know, (a), what it is for something to be *a* and, (b) what it is for something to be *F*. The person capable of a thought that *a* is *F* is then capable of deploying those distinct capacities over a range of all the other concepts she is in possession of. Thus, the ability to grasp the thought that *a* is *F* implies an ability to grasp the thoughts that, for instance, *a* is *G*, that *a* is *H*, and so on for each concept of property that someone has. Similarly, someone capable of a thought that *a* is *F*, must know what it is for something to be *F* which implies an ability of grasping the thoughts that *b* is *F* and *c* is *F* and so on for any object that someone can think of. A genuine thinker is someone who is capable of entertaining syntactically permissible[93] combinations of all the concepts in her possession. This ability of entertaining syntactically permissible combinations[94] of concepts is what counts as a conceptual competence and is something that animals are not capable of.[95] Or so it is argued. To sum up, if rationality – or operating with reasons – is tied

[92] See Gareth Evans, *The Varieties of Reference* (Oxford: Oxford University Press, 1982), p.100.

[93] Clearly, there will be many `thoughts` that won`t count as permissible. A permissible way of combining *John*, *Peter*, *tall* and *small* would be, for instance, *John* is *tall*, and *Peter* is *small*; the impermissible one: *John* is *Peter*, and *tall* is *small*.

[94] The generality constraint is usually given a `syntactic` reading and that seems to correspond to Evans` original conception of it. Myself, I don`t think that appeals to the generality constraint under the syntactic reading pack much punch against the possible claim that at least some animals are conceptually competent. There are several rather serious problems with the constraint. The problems are identified and discussed by Imogen Dickie, `The Generality of Particular Thought` *The Philosophical Quarterly*, 60, no. 240 (2010), pp.508-532. The force of the generality constraint in the context of alleged conceptual competence of higher animals can, however, be recovered if it is given a `semantic` reading, i.e. a reading which takes the constraint to be claiming that to grasp any concept at all, however elementary, one must already have a repertoire of concepts, a range of concepts that provide a semantic matrix within which each concept is intelligible. The generality constraint semantically interpreted would then be combined with a claim that animals simply do not have a repertoire or a range of concepts numerous and rich enough to sustain a grasp of even a single concept. Imogen Dickie (albeit in a different context) seems to me to be offering something like this semantic reading of the generality constraint when looking for ways to save the constraint from the problems she herself had identified (for more, see her paper mentioned above in this footnote).

[95] Some disagree. For instance, Peter Carruthers gives a convincing example of a `monkey that is familiar with an aged matriarch – call her `Elsa` - [that] might be incapable of thinking that Elsa is an infant` because `what could we possibly do that might induce the monkey to entertain such a thought (whether believing it, desiring it, or whatever)?` See his `Invertebrate concepts confront generality constraint (and win)`, in R. Lurz (ed.), *The Philosophy of Animal Minds* (Cambridge: Cambridge University Press, 2009), pp.89-107 (p.95).

with possession of language or/and conceptual competence then non-human animals cannot be understood as rational or intentional agents. Those committed to tying rationality with linguistic and/or conceptual competence will thus dismiss any employment of intentional explanations and intentional verbs in the case of non-human animals as a misleading (if useful) *façon de parler.*[96]

Some philosophers, on the other hand, will be committed to a *less* demanding account of *reasons* and/or *operating* with *reasons,* and this will allow them to attribute intentional action onto non-human animals. They might, for instance, refuse to tie rationality with conceptual competence[97]. With the requirement of the conceptual competence out of the equation, those philosophers will be looking for behavioural similarities between humans and a target non-human animal as a sufficient justification for attributing a corresponding intentional state. For instance, a chimpanzee trying to open a locked box with her favourite food inside will be seen as engaged in a rational problem-solving activity on the grounds of her demeanour (using and discarding one type of tool for another, scratching their head, etc.) being sufficiently similar to the behaviour of a rational human in the same situation. Another possible strategy would be to defend an account of conceptual competence that is `weak` enough to be applicable to non-human animals.[98]

Now how does the above said help us with the challenge of *deciding* having possibly nothing to do essentially with *weighing reasons* and/or *considering* things? Say you are committed to the demanding or - as I shall call it - a `heavy` account of reasons and operating with reasons. As such you will refuse the talk of non-human animals *deciding* anything because deciding is an intentional behaviour, which is something that, on this view, non-human animals are incapable of. You will not see any force in the appeal to newspaper articles that seem to take dogs as *deciding* (as you don`t accept the premise of a dog being able to decide anything) and you will dismiss the articles and their authors as having resorted to an anthropomorphic manner of speech. Consequently, you stick with the default position of understanding *deciding* as *weighing reasons* and/or *considering* things. If, on the other hand, your account of reasons and operating with them is suitably less demanding then you will very probably feel little or no discomfort when talking about apes (or dogs) as *deciding* things. You will be ok with that because although you accept that *deciding* is *weighing*

[96] Davidson puts employing intentional explanations in the case of animals on a par with employing them in the case of a heat-seeking missile. See Donald Davidson, *Subjective, Intersubjective, Objective* (Oxford: Oxford University Press, 2001), p.201.

[97] See for instance Hans-Johan Glock, `Animals, thoughts and concepts`, *Synthese*, 123, No. 1 (2000), pp.35-64.; Hans-Johan Glock, H.-J., `Can animals act for reasons?`, *Inquiry*, 52, no. 3 (2009), pp.232-254.

[98] That`s Peter Carruthers` strategy in his `Invertebrate concepts confront generality constraint (and win)`, in R. Lurz (ed.), *The Philosophy of Animal Minds* (Cambridge: Cambridge University Press, 2009) pp.89-107.

reasons, you have, at the same time, a `light` theory of reasons (and operating with them) that allows you to attribute intentionality onto non-human animals. Thus, you will not see any force in appeals to phrases of `dogs deciding things` either. Who will then?

There are at least two answers, a charitable one and an uncharitable one. Let`s get the uncharitable one out of the way first. Someone might be just rather inconsistent, mixing the `heavy` and `light` perspectives in one and the same context. Recall the challenge: People sometimes say that dogs *decide* things. We might not want to say, however, that dogs *weigh reasons* or *consider* things. That would suggest that *deciding* does not entail *weighing reasons* and/or *considering* things. But the problem arises only when you employ a `light` reading of *deciding* while, at the same time, employing a `heavy` one of *weighing reasons*. That`s clearly inconsistent though. You either employ the `heavy` reading in both cases - and that will not allow you to attribute a capacity for deciding to dogs – or you employ the `light` reading, in which case you won`t be able to conclude that *deciding* does not entail *weighing reasons*.[99]

The charitable answer is more interesting. There is an argument to the effect that deciding is not an intentional behaviour.[100] The conclusion of this argument can be employed as a premise in an argument behind the charitable answer. I will call the argument behind the charitable answer: *The argument from the unintentionality of deciding*. Intentional behaviour is – as mentioned above – a behaviour that is both motivated by *reasons* and involving an ability to, in some sense, *operate* with *reasons*. Thus, should it turn out that deciding is not an intentional behaviour we would be forced to conclude that it has nothing essentially to do with operating with reasons (or with weighing reasons). Consequently, even those committed to a `heavy` account of reasons couldn`t but attribute deciding to dogs.

There is an ambiguity here that needs to be exposed. Deciding can have something to do with reasons and reasoning in two senses. First, in virtue of it being an intentional behaviour. I will call this sense in which deciding has something to do with reasoning *an*

[99] It is worth mentioning that the account of freedom in agency that I will ultimately arrive at is compatible with both `heavy` and `light` accounts of reasons. The only difference resulting from a commitment to one of them concerns the scope of attributability of *freedom*. Those committed to `heavy` accounts will be able to attribute *freedom* to human agents *only*, while those committed to the `light` one will be able to – or will have to – allow attributions of *freedom* not only onto humans but onto (higher) non-human animals too (provided they allow for non-human animals to possess consciousness because the argument requires that any candidate eligible for attributions of *freedom* to be capable of having a *sense* of alternative). It should be noticed that the possible `light` reading of my account cannot be taken as an *unwelcomed* consequence of my argument because the consequence cannot be drawn by those who would find it *unwelcomed* – i.e. by those committed to the `heavy` account – while those who can draw it – i.e. those committed to the `light` account – will not find it *unwelcomed.*

[100] A concise summary and discussion of the argument can be found in Joshua Shepherd, `Deciding as Intentional Action: Control over Decisions`, *Australasian Journal of Philosophy*, 93, no. 2 (2015), pp.335-351.

intentional sense. Second, in virtue of its meaning. I will call this sense in which deciding has something to do with reasoning *a semantic sense*. To understand the distinction here, contrast *deciding* with *looking* (at someone). Both refer to intentional behaviour, thus some motivating reasons (with some elementary reasoning being involved) are implied. However, the two differ importantly with respect to their meaning. If a learner of English asks what *looking* (at someone) means, the answer will not mention any reasoning at all. If, on the other hand, the same question is asked about *deciding,* the answer will have to refer to *reasoning* (or some synonymous alternative) to capture the meaning.

Now, we could perhaps use this distinction to dismiss the charitable answer in the following way: *The argument from the unintentionality of deciding* targets only the *intentional sense* in which deciding has something to do with reasoning. It leaves the *semantic sense* untouched, and that`s all we need to be able to maintain that deciding has something essentially to do with reasoning. Unfortunately, the seemingly easy way to dismiss the charitable answer might not do. It could be argued that there is a relation of dependence between the *intentional sense* and the *semantic sense.* More specifically, it could be argued that *reasoning* in the *semantic sense* can be attributed *only* where it can be attributed in the *intentional sense*. It is difficult to disagree here. It seems to be clear that the presence of reasoning in the *semantic sense* entails its presence in the *intentional sense*. To deny this would be like denying that theorizing about language entails the existence of language. Consequently, the absence of reasoning in the *intentional sense* would imply its absence in the *semantic sense*. Or, in other words, once the conclusion of *the argument from unintentionality of deciding* is accepted, we won`t be able to take *deciding* as entailing *weighing reasons*.

There is a relatively easy way out though. Let`s have a look at the argument behind the claim that deciding is not an intentional action. The argument is presented and discussed by Joshua Shepherd in his `Deciding as Intentional Action: Control over Decisions`.[101] It can be summed up in the following way:

1. We ask: In virtue of what are decisions intentional?
2. A widely adopted answer:[102] Practical decisions are intentional in part because of the causal work of a mental state extrinsic to the decision itself – an intention to decide what to do.

[101] Joshua Shepherd, `Deciding as Intentional Action: Control over Decisions`, *Australasian Journal of Philosophy*, 93, no. 2 (2015), pp.335-351.

[102] And one given by Alfred Mele in his *Motivation and Agency* (Oxford: Oxford University Press, 2003), ch.9.

3. Typically, intentions that are relevant to intentional actions will, among others, *guide* actions.
4. The content of an intention must be *specific* enough to do the guiding.
5. An intention to decide what to do is essentially open-ended; i.e. it is not known *what* will be decided.
6. The open-endedness of an intention to decide what to do makes the intention intrinsically non-specific.
7. The intrinsic non-specificity of an intention to decide what to do makes it irrelevant to the (intentional) action. [3, 6]
8. Conclusion: The irrelevance of an intention to decide what to do regarding deciding what to do shows that deciding is not an intentional action.

It is not difficult to spot a problem in the argument. It is to be found in steps 5 and 6. It is claimed, in step 6, that an intention to decide what to do is intrinsically non-specific because of the open-endedness of such an intention. An intention to decide what to do is claimed (in step 5) to be open-ended on the grounds that it is not known what outcome will the intended deciding generate. It is hard to see, however, how not knowing in advance the outcome of the intended *deciding* – i.e. how the open-endedness of the intended *deciding* – makes the *intention* (to decide) non-specific in any relevant way. When I intend *to decide* (what to do), my intention is perfectly specific as long as I understand the notion of deciding. Thus, if I take deciding to be about weighing reasons then the content of my intention is sufficiently clear, robust and specific: *To determine what to do, my intention is to weigh reasons.* The open-endedness of deciding (and the resulting non-specificity) doesn`t transfer to intending. Compare an intention *to decide* what to do with an intention *to ask your friend* what to do. Just like in the case of intending *to decide* what to do, when I intend *to ask my friend* what to do, I have no idea what her advice will be. Yet it would be plainly wrong to claim that the very *intention* to ask my friend is therefore non-specific or not specific enough. *Asking one`s friend* is an action that simply is perfectly specific and any intention to perform it must be taken as similarly specific.

Let me briefly sum up here by getting back to the original question: Who will see any force in appeals to talk of `dogs deciding things` (in the context of settling down the issue of whether or not *deciding* entails *weighing* reasons)? The uncharitable answer points to those who mix `heavy` and `light` perspectives on what counts as acting for reasons in one and the same context. The charitable answer turns on something more sophisticated. It turns on *the argument from unintentionality of deciding.* A proponent of this argument will insist that the talk of dogs deciding things presents a serious challenge to the claim that *deciding* entails *weighing reasons.* She will insist so on the grounds that such a talk is correct, and that it is correct because it takes deciding as an unintentional action. An unintentional action is not motivated by reasons. Therefore, the proponent of the argument will conclude, if deciding is an unintentional action then *deciding* can`t have anything to do with *weighing reasons*. We have, however, seen that the central premise of *the argument from unintentionality of deciding* is untenable. The central premise claims that deciding is unintentional action. The premise won`t survive a closer scrutiny though. The argument behind the premise confuses the open-endedness and non-specificity of deciding with the open-endedness and non-specificity of an intention (to decide). As such, it fails to show that deciding is not an intentional action. This collapses *the argument from unintentionality of deciding.* Thus, the charitable answer – i.e. the answer that employs *the argument from unintentionality of deciding* – will have no force against the claim that *deciding* entails *weighing reasons*.

2.5.3 Deliberation: robust and derivative

There is one more issue I need to touch upon before summing up the discussion of (iv), that is, the discussion of the claim that *choosing* entails *deliberation.* It might be objected that we commonly take people as choosing and deciding things even in those everyday situations where we would find it rather unnatural to say they are weighing reasons or considering and comparing features. True, it will be conceded, there are clear cases of choosing and deciding that involve rich and robust instantiations of weighing reasons and/or considering and comparing features. A young couple choosing a house to buy will exercise a lot of weighing reasons and/or considering and comparing features. There are, on the other hand, similarly clear cases of choosing and deciding that happen either too fast or in an `autopilot mode` where it feels wrong to say that weighing reasons or considering and comparing features has taken place. During a ski race an alpine skier *chooses* an optimal trajectory between the gates. The choosing happens in a fraction of a second and there is no time to weigh reasons or consider and compare features. The skier doesn`t deliberate. Or, driving home from work you *decide* to stop at a café to get a sandwich, which is what you nearly always do on your way home. You were far from weighing any reasons or considering and comparing anything. You didn`t deliberate. You were in an `autopilot mode`. Now does this show that *choosing* doesn`t entail *deliberation* after all?

I don`t think it does. True, choosing and/or deciding in an `autopilot mode` doesn`t involve weighing reasons in a robust sense; it does, however, involve it in a sense that is *derivative* from the robust one. Presumably, one will switch to and go into an `autopilot mode` – i.e. turning off any weighing of reasons when choosing and/or deciding – only if the relevant action is a token of a type action whose reasons had been weighed and endorsed before and the action turned out to be a success. Applied to the scenario of a driver on his way home from work popping in the same café to get a sandwich, it is quite natural to expect that there is some history to this. Perhaps something like this. First week in a new job. Driving home after work; getting hungry and thinking and weighing reasons regarding what to do about this. Shall I hold on and eat at home? I could save some money plus eat in the company of my wife. Or shall I look for a café somewhere on the way where I could have a sandwich and a cup of tea? I could have a little moment just with myself and my thoughts to reflect on the new job and my new colleagues. I deliberate and choose to get a sandwich in a nearest café. The café turns out to be a cosy little place with excellent choice of sandwiches, loose leaf teas and relaxing music. Driving home from work next day I again get hungry and choose to get a sandwich in that same café. There is some deliberation involved but far less of it then yesterday. Dining with my wife still has the same appeal and I also believe – as I did yesterday - that it`s one of those little things that contribute to a happy relationship but the café and the sandwiches and the relaxing music were so enjoyable the day before that today this alternative has much more force in my reasoning. I also really enjoyed the time and opportunity to reflect on various issues going on in my life. This time, the deliberation is brief and `shallow` (or more `shallow`). The alternative to the sandwich-in-the-café one – i.e. driving straight home and having dinner with my wife - doesn`t really have much pull any more. After a few days, going to the café becomes part of my routine. I don`t deliberate. I switch to an `autopilot mode`. Not only do I not weigh reasons anymore, I am able to think about various unrelated things while *choosing* to stop by at that café. Yet it is correct, I believe, to insist that my making a decision to stop by the café involves weighing reasons. It does so not only in the historical sense described above - that would perhaps be felt as too weak – but it is there in the *real* time (even if only in the derivative way and not in the phenomenologically rich way as in the case of the young couple deliberating about the purchase of a new house). The evidence for this is that it generally makes sense to ask a human agent who made a decision in an `autopilot mode` for reasons which carried weight in that decision. Typically, the agent will *recover*, not *perform*, the reasoning behind the decision. That is, the agent will bring forth what is already there and involved in the relevant choosing and deciding.

As for the cases when choosing and/or deciding happens too fast to involve weighing reasons in some phenomenologically robust sense – such as that of a skier choosing an optimal trajectory between the gates – most of them involve deliberation in a similarly derivative way as the `autopilot mode` cases. The skier processing the features of the race course and choosing the optimal trajectory is in an `autopilot mode` too. The difference is that – compared to the driver`s `autopilot` mode – her `autopilot` mode is just a much faster processing one. The sense in which the skier`s choosing the optimal trajectory involves – in a derivative sense – weighing reasons is something like this: Throughout many years of intensive training the skier`s coach explained to her repeatedly various physical relations among the speed, the gravitational pull, the width of the curve, the angles of approaching the gate, the point of the start of a turn, and so on. Before each race, the skier will get thoroughly acquainted with the race course. This will involve at least one (usually two) test rides and a slow walk through all the gates of the race course. When getting acquainted with the race course (especially when going through it on foot) the skier will apply her knowledge of the physics of optimal trajectory to determine the best possible trajectory to get through each gate. The application of her knowledge can be perfectly naturally described as weighing reasons and/or considering and comparing features. Then she memorizes the trajectory she has chosen for each section of the race course. It is correct to say that during the race she *chooses* the optimal trajectory - even though no weighing of reasons at or immediately prior to the *choosing* is taking place – solely on the grounds that the relevant deliberation is involved in the act of choosing in a derivative sense.[103] The derivative sense is strong enough to sustain the target entailment.[104] And here too we can test the claim by asking the skier why she chose this and not some other trajectory. She will *recover*, not *perform*, the relevant reasoning in her reply.[105]

[103] An appeal to derivative presence of reasoning will play an important role in the overall plausibility of my account of freedom in agency. I will tie attributions of freedom with an exercise of reasoning. The derivative notion of reasoning will allow me to attribute freedom to agents in a variety of everyday situations where such an attribution feels very natural and plausible even though the agents are clearly not engaged in any robust exercise of reasoning.

[104] You might disagree. You might insist that this kind of historical reconstruction of how deliberation is present in the agent`s choosing doesn`t show that it is there essentially. Consider, however, what happens in scenarios where it is clearly indicated that no historical and no co-temporal deliberation took and/or could have taken place: *Walking in the forest I suddenly hear a loud crack somewhere above me. I look up and this huge branch is off and about to crash on my head. I lunge forward, head first, landing on my belly at the exact moment when the huge branch hits the exact spot I was standing on just a fraction of a second ago.* Now, try to insert `choose to` in between `I` and `lunge` in the third sentence of this little story. It just feels completely wrong to say `I choose to lunge forward…`, and the reason is that the scenario makes it clear that no deliberation whatsoever – derivative or non-derivative – took, or could have taken, place.

[105] There will be grey zones here. Perhaps the skier`s coach has never explained to her the physics behind choosing an optimal trajectory. Perhaps he thought she was too dumb to understand it. Instead he always uses a 3D computer modelling of the race course, determines the best trajectory himself and then plays it to the skier in her 3D goggles. She memorizes and test-rides it. I suspect that scenarios like this will turn divisive regarding the applicability of the notion of choosing. Some might feel it ok to describe the skier as *choosing* the trajectory on the grounds that her choice is a result of deliberation although it is a deliberation performed not by her but by her coach. It could be seen as a case of a weaker – sufficient nonetheless - derivative presence of deliberation, a case where the presence of deliberation derives not only historically but also from a mental process of a separate individual. There is a helpful and well-known analogy. In philosophical semantics a speaker is taken as having a robust grasp of a concept just in virtue of her deference to experts regarding the conditions of applicability of that concept. Similarly, some might be fine with the talk of *choosing* if it involves the skier`s deference to the deliberation of her coach. Others will find this too weak and refuse to talk here about *choosing* at all. I am inclined to side with those who find it unnatural to describe the dumb skier as *choosing* although I can see why others might be fine with such a description. The important bit to notice here is that both sides will theorize under the assumption that *choosing* has something essentially to do with *deliberation*.

2.5.4 Subconclusion

Let me briefly sum up and conclude. The claim was: *choosing* entails *deliberation.* In our ordinary language we understand *choosing* as *deciding what one wants* or/and *thinking about which one it is that one wants*, and *deliberation* as *weighing reasons* or/and *considering* and *comparing* features. The phrase *thinking about which one it is that one wants* seems to say the same as the phrase *weighing reasons or/and considering and comparing features* in the context of determining the preferable option. With *choosing* understood as *thinking about which one it is that one wants*, the exposition of the entailment – that *choosing* entails *deliberation* - is, therefore, very quick and straightforward. It gets somewhat less straightforward when *choosing* is analysed in terms of *deciding what one wants*. There are two challenges here.

First, there are contexts within which *deciding* seems to be semantically divorced from *weighing reasons* and/or *considering* and *comparing* features. Such contexts constitute a challenge as they would, if not neutralized, prevent me from arriving at my target claim that *choosing* entails *deliberation*. However, a closer scrutiny of the contexts reveals an inconsistency and/or confusion. With the inconsistency removed and/or the confusion clarified, the first challenge is neutralized and we fall back on the natural reading of *deciding what one wants* as involving *weighing reasons* and/or *considering* and *comparing* features.

Second, there is an argument from the unintentionality of deciding that, if successful, severs the conceptual tie between *deciding* and *weighing reasons*. The argument rests crucially on the claim *that deciding is not an intentional action.* I refuted the claim, which collapsed the argument behind the second challenge.

I then proceeded to discuss a possible objection concerning the plausibility of the claim that any choosing or deciding involves deliberation. We commonly describe agents as *choosing* and/or *deciding* even when it doesn`t seem to be the case that the agents are – simultaneously or shortly before - engaged in any weighing reasons and/or considering and comparing features. I argued that the involvement of deliberation in choosing and/or deciding doesn`t have to be simultaneous or immediately preceding. Often, our choosing and/or deciding is grounded in a deliberation that was performed a long time ago. The presence of such a past deliberation in a present act of choosing and/or deciding is derivative but robust enough to sustain the target entailment. At this point nothing prevents us from concluding (iv): *Choosing* entails *deliberation*.

2.6 *Deliberation* entails *an exercise in reasoning*.

This one is clear and uncontroversial. Above, we take *deliberation* to mean *weighing reasons* (and/or *considering* and *comparing* features of available options). That`s what deliberation simply is. Similarly clear and uncontroversial is taking *weighing reasons* as a case of *an exercise in reasoning*; *weighing reasons* just is *an exercise in* reasoning. It follows then that: *deliberation* **is** *an exercise in reasoning*. Here I shall assume something, again, rather uncontroversial which is that an identity relation is a variety of an entailment relation,[106] and conclude that *deliberation* **entails** *an exercise in reasoning.*[107]

The reader might concede that there is, indeed, not much to object against here but, at the same time, feel somewhat cheated regarding the informativeness of the target claim. *Reasoning* and the related notion of *rationality* are philosophically loaded notions and the reader`s grasp of the entailment in the target claim will necessarily be rather tentative as long as its members remain theoretically `unprocessed`. I will say more about *reasoning* and *rationality* below, when discussing and defending the plausibility of my account of freedom in agency. Here my task was much simpler: to make the reader pause for a while and agree that *deliberation* has essentially something to do with *reasoning*.

2.7 *Freedom in agency* entails *an exercise in reasoning*.

This is the conclusion I wished to arrive at. It follows from (i)–(v) in virtue of entailment being a transitive relation.[108] Or does it? The following could be objected. In steps (ii) and (iii), we talk about *having a sense of* an alternative and *having a sense of* being able to choose while the following three steps, (iv), (v) and the conclusion itself, present claims that lack the phenomenological constraint (of *having a sense of*). Shouldn`t those last three steps also

[106] Anderson and Belnap call the variety of entailment that an identity relation is a *tautological* entailment. See Alan Ross Anderson and Nuel D. Belnap, Jr., `Tautological Entailments`, *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 13, no. 1/2 (1962), pp.9–24. I shall ignore here certain complications stemming from the fact that identity is a symmetrical relation while entailment is not. For a discussion of some paradoxes that this fact leads to, see C. Lewy, `Entailment and Propositional Identity`, *Proceedings of the Aristotelian Society*, New Series, 64 (1963 – 1964), pp.107–122. I feel comfortable with ignoring the complications and paradoxes discussed by Lewy because, ultimately, they must be taken as a challenge to deal with and not as a proof that identity doesn`t imply entailment. To accept them as such a proof would be way too disruptive for our common way of reasoning about virtually anything.

[107] A reminder: The identity relation and the entailment relation referred to here are not of the strong kind, i.e. of the kind that implies (conceptual) necessity. Above (pp. 24-25), I have briefly touched upon the unviability of the notion of conceptual necessity. For the lack of a better option I stick with the notion of entailment but the reader is asked to remember that I use the notion as implying a strong modal tie to truth and not a necessity. The same goes for the notion of identity.

[108] Again, there are some complications related to the notion of transitivity of entailment. See D. L. C. Maclachlan, `The Pure Hypothetical Syllogism and Entailment`, *The Philosophical Quarterly*, 20, no. 78 (1970), pp.26–40. Maclachlan argues against the transitivity of entailment. But see Frank Jackson who counters convincingly Maclachlan`s argument in his `The Transitivity of Entailment`, *The Philosophical Quarterly,* 20, no. 81 (1970), pp.385–388. I am not too disturbed by the complications. Even if Maclachlan is right about entailment not being a transitive relation, it won`t affect my argument here. Maclachlan`s treatment of entailment takes it as implying a necessary relation. This ultimately leads to paradoxes for reasons that Timothy Williamson appeals to in his rejection of the notions of *conceptual necessity* and *conceptual truth*. I don`t, however, understand the notion of entailment here as implying necessity - just a strong modal tie to truth – thus Maclachlan`s argument doesn`t have much force against my position.

be embedded within the phrase *having a sense of*? Thus, instead of `*choosing* entails *deliberation*`, (iv) would read as `*having a sense of choosing* entails *having a sense of deliberation*`; and instead of `*deliberation* entails *an exercise in reasoning*`, (v) would read as `*having a sense of deliberation* entails *having a sense of an exercise in reasoning*`; and, finally, instead of `*freedom in agency* entails *an exercise in reasoning*`, the conclusion would then read as `*freedom in agency* entails *having a sense of an exercise in reasoning*`. And that`s a somewhat different conclusion.

There a few ways of answering the objection. The simplest one is this. Regarding (iv) and (v), it would be unnecessarily cumbersome to embed the entailment in the phrase *having a sense of*. It is uncontroversial that *choosing* entails *deliberation* only if choosing involves essentially deliberation. Then, of course, having a sense of choosing will involve essentially having a sense of deliberation. And this will allow us to say that *having a sense of choosing* entails *having a sense of deliberation*. Exactly the same logic applies to the claim that *deliberation* entails *an exercise in reasoning.* This claim analogously implies that once one has a sense of deliberation then one must have a sense of an exercise of reasoning, which allows for the conclusion that *having a sense of deliberation* entails *having a sense of an exercise in reasoning*. That is, once the truth of (iv) and (v) has been established, the truth of their phenomenologically constrained variants too gets established. Thus, there is no need to be explicit.

All this might be conceded. The core of the objection remains unaffected though. The above seems to confirm the suspicion that the valid conclusion of the argument should be: *freedom in agency* entails *having a sense of an exercise in reasoning.* That is, strictly speaking, correct. However, it becomes an issue only if it is presumed that one can have a sense of an exercise in reasoning without there being an actual exercise in reasoning. And that seems inconceivable. True, one can mistake one`s sense of something else for a sense of an exercise in reasoning, or the very exercise of reasoning that one has a sense of can be deeply flawed (i.e. involving conceptual confusions or logical errors). Neither shows, however, that one can have a sense of an exercise in reasoning without actually exercising her reasoning. It follows then that the claim that *freedom in agency* entails *having a sense of an exercise in reasoning* can be read as implying that *freedom in agency* entails *an exercise in reasoning*.

The argument (i) – (vi) is meant to back up (C).

(C): Freedom is attributable only to agents who exercise (practical) reasoning.

The conclusion of the argument – *freedom in agency* entails *an exercise in reasoning* – gets reformulated into (C) in the following straightforward way. It is uncontroversial that (A ⊨

B) → (¬B → ¬A), i.e. it is uncontroversial that from `pain entails sentience` it follows that `if there is no sentience, there is no pain`. In terms of *attributability* it says that `pain` can be *attributed* only where `sentience` is instantiated. Similarly, the entailment in (v) implies that if there is no reasoning then there is no freedom. And in terms of *attributability* we get to a variation of (C): `freedom` can be *attributed* only where `reasoning` is instantiated. I take it that enough has been said to establish the truth of (C).[109]

Let me make two preliminary remarks here to prepare the ground before moving on to the discussion in the next chapter. First, (C) identifies a *necessary* condition, not a *sufficient* one. It will be the task of the next chapter and the discussion there to show in what sense (C) can be understood as capturing the sufficient condition of attributability of freedom. Second, I wish to draw the reader`s attention to where we are, dialectically speaking, regarding the progress we have made in solving the problem of free will/freedom in agency. In our pre-theoretical understanding, freedom is often taken to have something to do with a lack of constraints. (C), however, gives us a rather contrary picture of freedom. It ties attributions of freedom to an exercise of reasoning. Reasoning is a rule-following activity, that is, an activity that is *constrained* by rules. This implies that freedom too is, essentially, something *constrained*. Now, the problem of free will/freedom in agency is to figure out how freedom can be instantiated in the world that is fundamentally *constrained* by causal determinism. With (C), however, it appears that instead of somehow weakening the *constraining* grip of causal determinism, we have identified another *constraining* layer: the rule-following exercise in reasoning. Thus (C) - far from, at this point, looking promising in some way - could be seen as worsening the incompatibility dilemma of the problem of free will/freedom in agency. There is a way of turning things around though. In the next chapter, I will present a thought experiment that will allow us to look at (C) from a novel perspective; a perspective within which it will be possible to interpret (C) as a claim about a *sufficient* condition of attributability of freedom.

[109] Let me recap an important point here: (C) is *not* a stipulative or normative claim. It is not, implicitly, saying that we *should* attribute freedom only to reasoning agents (it, perhaps, does so only on a meta-level in virtue of the rational being intrinsically normative). (C) is a metaphysical claim. It tells us that the nature of freedom (in agency) is such that it can get instantiated only within the domain of a reasoning agent.

Chapter 3: From reasoning to freedom

3.1 The introduction

The aim of this chapter is to show how the conclusion that we arrived at in Chapter 2 can be understood as identifying a sufficient – and not just a necessary – condition of attributability of freedom onto an agent. The central role in this is played by a thought experiment presented below.

3.2 The thought experiment

Two things need to be introduced first.

1. A basic, uncontroversial picture of agency:

Input (the relevant causal history) → (conceptual and rational) **Processing → Output** (action)

I will call this an IPO picture of agency. The thought experiment as presented below will provide an illuminating context should the reader find this picture of agency unclear.

2. A concept of veto power.

The concept of veto power is a minimalistic version of the default concept of free will. *Free will* as *a veto power* refers to the power to do A or to refrain from doing A. The default understanding of free will – the understanding that problematizes *freedom* in the causally determined world – takes it to be *a power to do otherwise*. Clearly, veto power as a power to refrain from doing something is a case of a power to do otherwise.

The thought experiment:

The setting.

Agents A1 and A2. The agents are identical in all respects except for the following: A1 has no veto power while A2 has it. Scenarios S1 and S2. The scenarios are identical in all respects except for the following: S1 contains A1 while S2 contains A2. Both S1 and S2 are causally determined in the following way: The Input-Processing-Output, (IPO), triad in S1 and S2 will be identical as long as A2 (in S2) doesn`t exercise her veto power.[110] A1 and A2 have

[110] There is an important sense in which the world can be understood as causally determined even if it is inhabited by agents capable of (and occasionally performing) a veto over a causal chain running through their agency. There will still be lawlike causal relations out there in such a world. Our knowledge, our theories about such a world will be the same or very similar to the ones we possess, the only difference being that formulations of such knowledge will start with an implicit or explicit

an identical sense of available alternatives, i.e. they have phenomenologically indistinguishable acquaintance with choosing *freely*. Neither of the two knows whether their experience is of a real capacity or a mere illusion.

The plot.

An identical causal history (an identical Input) has brought both A1 and A2[111] (in their respective scenarios) into a room with a desire to hang a favourite painting on the wall. On the table in front of them, there is a nail and a hammer. If A2 doesn`t exercise her veto power, the following will happen. Both A1 and A2 process (they go through the Process stage of the IPO picture of agency) the Input (the perception of nail and hammer, the knowledge of what these objects are for, the desire to hang the painting and so on) and act (generate Output): both grab the hammer, drive the nail into the wall and hang the picture. That is, the (IPO) in S1 and S2 will be identical.

The twist.

A2 is a veto power capable agent. The veto power makes agent A2 *free*[112] in a very strong sense; perhaps the strongest conceivable one.[113] [114] She could exercise her veto power, block

qualification: `Outside of the agential interference, the world follows this and that set of laws...`. I think this qualification might feel less unpalatable if we recall that all the laws of science are formulated within an (implicit) set of analogous *ceteris paribus* qualifications.

[111] Assuming A2 has, so far, never exercised her veto power.

[112] Freedom in will/agency is, uncontroversially, understood as the power to do otherwise. This, presumably, wouldn`t be denied by even Frankfurt and his followers who argue that moral responsibility (and freedom too, in their understanding) is attributable to agents even in contexts when they are incapable of doing otherwise. Their concept of freedom – being the weaker one – doesn`t exclude the strong one: freedom as power to do otherwise. Clearly, the veto power – the power to block or refrain from blocking a relevant causal chain - is a case of power to do otherwise.

[113] It could be objected that there is a stronger notion of freedom than the veto-power one. Such a stronger notion would be the causa-sui one. The causa-sui freedom would be a freedom where the agent is capable of originating a brand new causal chain in a sort of positive way without any prior causal determination. The veto-power freedom overlaps conceptually with the causa-sui one to the extent to which they both are causal events disconnected from any causal history. To this extent the veto-power freedom is a variety of a causa-sui action. The veto-power freedom differs from the causa-sui one in that the former doesn`t initiate a brand new causal chain in any sort of a positive way. It works in a purely negative way; just blocking a relevant causal chain thus allowing a different existing causal chain to (potentially) take over and be realized in action. As such the veto-power freedom seems to be, relatively, an easily conceivable one. The causa-sui freedom, on the other hand, borders on inconceivability as there seems to be much more mystery involved in originating a brand new causal chain than there is in just blocking one.

[114] Some naturalistically oriented philosophers tend to look at thought experiments and their epistemological value with suspicion. Those philosophers might find the thought experiment more palatable once reminded of a well-known scientific experiment that is, in relevant respects, analogous to the A2/S2 (part of the) thought experiment. The scientific experiment (or, to be more precise, a series of them) was conducted by the neurologist Benjamin Libet and is described in his paper `Do We Have Free Will?`, *Journal of Consciousness Studies*, 6, no. 8-9 (1999), pp.47-57. Libet found out that any awareness of an intention to act is always preceded by a relevant neural activity. That makes all our decisions to act (remember: deciding is an intentional process) predetermined by neural processes in the brain. At the same time, and quite surprisingly, Libet also found out that the subjects

the causal chain and refrain from generating the Output. Now, will she *choose* to do so? And in case she does, will that count as an exercise of freedom (in agency)? I claim `no` is the correct answer to both questions.

3.3 A conceptual interpretation of the thought experiment: *Reasoning in agency* as entailing *freedom in agency*

Two questions to be answered:

Q1: Why will agent A2 *choose* not to interfere (i.e. use her veto power to block the relevant causal chain)?

Q2: How does answering (Q1) help us recover *freedom* in S1 scenarios, i.e. in scenarios that assume a full causal determination of the world and an impossibility of a veto power?

Regarding (Q1): In chapter two I argued towards the conclusion that *freedom* is attributable only to agents who exercise practical reasoning. The agency is constituted by actions. An action that is a result of an exercise of practical reasoning is a *rational* action. *Freedom* is attributable only where rational action is instantiated, that is, an irrational action is not to be taken as an exercise of an agency that sustains attributability of *freedom*. I will call the agency that sustains attributions of *freedom*: a freedom sustaining agency or an FS-agency. And because an irrational action is not constitutive of the FS-agency I will want to say that: *In an irrational action the FS-agency is suspended.*

In the S1/S2 scenario, grabbing the hammer, driving the nail into the wall and hanging up the painting is *the* rational thing to do. Not doing *this* amounts to a collapse of the FS-agency. If A2 interferes and blocks the causal chain, she simultaneously suspends her FS-agency in the very act of interfering. Claiming that the *agent* A2 could have refrained – as in `could have *chosen* to refrain` - from performing the rational act reveals a confusion regarding the concept of agency and its relation to freedom. An irrational *agent* - that is, a non-FS-agent - cannot be attributed *freedom* and therefore, as such, can`t be said to be *choosing* anything. The conceptual logic behind the question whether the agent A2 will *choose* to block the causal chain is incompatible with an affirmative answer.

of the experiment were able to *block* the actualization in action of the relevant neural event; i.e. he noticed that the subjects had a *veto power* over the realization of the causal chains started by the subject`s brain. During the experiment(s) Libet wasn`t able to identify any relevant neural activity that had determined the exercise of veto power. Clearly, Libet`s inability to identify neural causes of subject`s vetoing a particular causal chain doesn`t mean that there are no such neural causes. Perhaps a different kind of experiment would have to be designed to identify them. The point here is that the philosophically uncorrupted mind of a scientist like Benjamin Libet doesn`t feel uncomfortable with conceiving of agents whose actions are causally *determined* and who are, at the same time, able to *block* the relevant causal chains.

Regarding (Q2): The answer given to (Q1) yields an intriguing result. It allows us to generalize something like the following.

> R: Whenever we act rationally (that is: whenever we exercise our FS-agency), our actions (that is: the actions of causally determined agents) are the *same* actions we would *choose* were we agents possessing the veto power over the relevant causal chains, i.e. were we *free* in a robust metaphysical sense.

This *sameness* of the actions constitutes a sense in which acting rationally – regardless of the fact that the rational processes themselves supervene on causally determined brain processes and are, moreover, essentially constrained in virtue of being rule-following processes – is *coextensive* with exercising agential freedom in the way that involves the ability to do otherwise.

I submit that what we have arrived at here is a robust compatibilist solution to the problem of free will. For those of us unable to ignore the force of the argument behind causal determinism, there is quite a lot of theoretical comfort to be drawn from knowing that whenever our action is rational it is exactly the same action we would have chosen in the same situation had we been free in the robust metaphysical sense. Bear in mind that the theoretical comfort to be drawn here will join forces with the everyday and inescapable experience of the phenomenology of choosing. This everyday and inescapable experience of the phenomenology of choosing is extremely rich and deeply constitutive of how we see and understand ourselves - it is so strong that many of us will refuse to really accept the truth of causal determinism with all its implications anyway. It is in this context that the robustness of my compatibilist solution should be assessed. The solution doesn`t really need to deliver a theoretical knock-out blow on its own because its target – the claim of causal determinism – is already seriously weakened by the fact of its non-theoretical implausibility and/or practical non-liveability.[115] Still, one might wonder whether a more robust or, in some sense, stronger claim could be squeezed out of the proposed compatibilist solution.

3.4 An existential interpretation of the thought experiment: Free agents exist

A specialist in the problem of free will might be perfectly happy with (R). She understands the parameters of the problem, that is, the dilemma constituted by the two *prima facie* incompatible claims: we are free agents and the world is causally determined. She – the

[115] The talk of practical non-liveability of causal determinism is an allusion to P. F. Strawson`s view that, roughly, it is practically impossible to give up treating others and ourselves as free and responsible agents. See Peter F. Strawson, `Freedom and Resentment`, in John M. Fisher and Mark Ravizza (eds), *Perspectives on moral responsibility* (Ithaca: Cornell University Press, 1993), pp.45-66. This view of Strawson`s will be discussed in more detail in Chapter 4.

specialist – accepts the truth of the individual claims constituting the dilemma and what interests her is whether there could be a way of thinking about the claims such that it would remove the *prima facie* incompatibility.[116] (R) represents the result of an attempt at such a removal of the *prima facie* incompatibility and the specialist will move on to another philosophical problem if she deems the attempt successful.

A non-specialist might feel cheated though. A non-specialist simply wants to know whether we *are* free or whether this can`t be the case because of the world being causally determined.[117] Giving (R) - that is: Whenever you act rationally your action is identical to the action that you would have chosen had you been really free – as an answer to the straightforward question of a non-specialist (Are we free agents?) will feel like a sophistic(ated) way of avoiding the question. Alternatively, the non-specialist might take (R) as entailing a negative answer to his question due to the unreal conditional that the formulation contains. Arguably, saying `you would have chosen [it], had you been really free` implies that you are, in fact, *not* free. And that`s not how I wish (R) to be read. I could perhaps dismiss the non-specialist on the grounds that his is a different question from that of mine. At the same time, it would feel rewarding to be able to answer exactly this question. It is, after all, this question that the majority of people out there associate with the problem of free will. And even a specialist will surely be interested in finding out whether (R) can be translated into an existential claim about free agency.

So, can (R) be shown to entail, imply or mean the same as the claim that: `We are free agents`? I believe it can. Below, I will formulate an argument that shows how. I will proceed in the following way. First, I will give a brief outline of the argument. Second, a discussion of the individual steps will follow.

The outline:

1. (R), and the related discussion, implies that an action being rational suffices to describe the action as `free`; that is: if an action is *rational*, it is *free*.

[116] `The Problem of Free Will` label gets attached to various positions or theoretical engagements related to free will/freedom in agency. Thus, you can have a specialist – such as Ted Honderich, for instance – who will not be interested in the problem of the *prima facie* incompatibility because he doesn`t think that we are free agents. His writings on free will focus on various implications of – what he believes to be – the fact of us being agents who lack free will (see for instance, Ted Honderich *How free are you?* (Oxford: Oxford University Press, 1993). So, there are specialist in the problem of free will who will not really be impressed by (R), but that`s mainly because they are interested in a different issue. I believe that the very core of the problem of free will is the issue of the *prima facie* incompatibility. All the other issues or problems that fall under `The Problem of Free Will` label derive from this core problem.

[117] There is a number of popular books that take on the question of free will. Whatever answer they give they, typically understand and frame the question as an existential one, i.e. as a question about whether or not metaphysical freedom and/or free agents exist in the causally determined world. A god example of such a popular book is, for instance, Sam Harris, *Free Will* (New York: Free Press, 2012).

2. The Principle of Undeniability of the Rational: it cannot be denied that a rational action exists.
3. Free action exists. [1, 2]
4. It is an empirical fact that people, at least sometimes, act rationally (in the sense that they generate rational actions).
5. An agent that generates rational action is herself describable as \`rational\` if a condition *c* is met.
6. It is an empirical fact that agents often give explanations of their rational actions such that it strongly suggests that condition *c* is, at least sometimes, met.
7. Rational agents exist. [4, 5, 6]
8. For an agent, being rational suffices to be describable as \`free\`; that is: if an agent is (being) rational, she is free.
9. Free agents exist. [7, 8]

Discussion of the individual steps:

(1): (R), and the related discussion, implies that an action being rational suffices to describe the action as \`free\`; that is: if an action is *rational*, it is *free*.

(R) says: Whenever we act rationally (that is: whenever we exercise our FS-agency), our actions (that is: the actions of causally determined agents) are the *same* actions we would *choose* were we agents possessing the veto power over the relevant causal chains, i.e. were we *free* in a robust metaphysical sense.

Recall how we got to the *sameness* of the actions in (R). First, in Chapter 2, we concluded that only an agent who exercises practical reasoning can be described as free. Regarding actions, this conclusion translates as: only actions that are a result of practical reasoning (of an agent) can be described as free. What we have here is a *necessary* condition of attributability of *freedom* onto agents and actions. The thought experiment has demonstrated that the possession of a veto power plays no role in the agent\`s *choosing* her action. The logic behind the conceptual relation among *choosing*, *rationality*, and *freedom* won\`t permit describing an *irrational* action (and its agent) as *free*. Thus, both the agent with her veto power switched on and the agent with her veto power switched off will *choose* the *same* rational action because that\`s the only one that can be *freely chosen*. The whole point in other words: it is uncontroversial that a vast majority (if not all) of philosophers would agree that an agent who possesses a veto power is *free* in a robust, metaphysical sense. The actions of such a metaphysically free agent are, of course, free (as long as they result from an exercise of practical reasoning). That much is clear. The thought experiment shows, however, that a metaphysically free agent cannot veto a rational action if she wishes to stay in the realm of freedom. Thus, once an agent is on the track of practical reasoning leading towards an action, the possession of veto power becomes irrelevant.[118] At the same time, as everyone agrees, the actions of a metaphysically free agent (who acts rationally) are free. Thus, there can be only one conclusion: If an action is *rational*, it is *free*.

(2): The Principle of Undeniability of the Rational: it cannot be denied that a rational action exists.

The basic idea here is this: denying is a rational activity; thus any attempt to deny that a rational action exists can be done only through generating a rational action – an action of denying – which is, clearly, self-defeating. Any attempt at denying the rational confirms it.

It seems to me that the only way to attack the basic idea here is to deny that denying is or must be a *rational* activity. There certainly are contexts where it is correct to talk of *denying* even though the action so described is not rational in any obvious sense or might even be outright irrational. There is a common usage of the verb `to deny` that simply means the same as saying `no`. A journalist will report on a politician saying `no` to a question as that politician *denying* a particular claim, accusation or interpretation. Simply saying `no` doesn`t have to involve anything that could be describable as a rational action while, at the same time, being correctly interpretable as *denying*. Now, clearly, there is nothing impossible or even mildly difficult in denying that a rational action exists if one takes *denying* to mean *saying `no`* (to a question, a claim or a proposition). However, and equally clearly, that is not how *denying* in (2) is meant to be, or needs to be, understood. In the market of philosophical claims any *denying* in the sense of simply *saying `no`* has absolutely no force. In the market of philosophical claims, any *denying* must be grounded in a rational argument to have any force. The success of a *denial* – that is, its impact on a philosophical argument or claim – will correlate positively with the force of the argument that the *denial* is grounded in. (2), of course, is a philosophical claim and must be understood as such.

There is a complication here that needs to be dealt with. It could be argued that the fact of the *undeniability* of the existence of something doesn`t entail the *existence* of that something. Thus, it is possible and rational to insist that, for instance, it is undeniable (in a sense explained below) that: yellow crocodiles that can beat Gary Kasparov in the game of

[118] The possession of a veto power is irrelevant even before getting on the track of practical reasoning preceding an action because exercising a veto power makes sense *only* in the context of choosing; and choosing is something that already involves essentially an exercise of practical reasoning.

chess exist. How could that be undeniable? Well, try to deny that. It is notoriously difficult (if at all possible) to conclusively deny any existential claim about an empirical entity or event.[119] It is possible to generate an infinite number of existential claims that are undeniable in the sense of them being impossible to deny conclusively. However, it would be absurd to commit to the truth of those claims on the grounds that we are unable to deny them. Therefore, the usual practice here is to place the burden of proof on those who make the existential claim and not on those who deny it. Applied to (2), the objection would then be:

> O: One cannot deny (conclusively) that a rational action exists, i.e. it is *undeniable* that a rational action exists, but it doesn`t follow that it is true that a rational action exists.

(O) is then accompanied by a reminder that the burden of proof is on those making an existential claim, which, in our case, is the claim that a rational action exists. In other words, (3) doesn`t, strictly speaking, imply that rational action exists; and an additional argument is needed.

Now, this additional argument for the existence of a rational action is relatively simple. Let`s compare our two existential claims:

> C1: Yellow crocodiles that can beat Gary Kasparov in the game of chess exist.
>
> C2: Rational actions exist.

(C1) is an existential claim that refers to empirical objects (yellow crocodiles) and an empirical event (beating Gary Kasparov in the game of chess). There is nothing that we know about the world that would indicate that, (a), there are yellow crocodiles, (b), that crocodiles of any colour could beat Gary Kasparov in the game of chess. On the contrary, all we know about crocodiles generally, and about Gary Kasparov and the game of chess suggests strongly that (C1) is not true. (C2), on the other hand, seems to be a default position. It is an empirical fact that a concept of rational action exists. True, there will be competing accounts of what counts as a `rational` action but that is not a problem because claiming that a rational action exists doesn`t commit us to a particular account. It is also an empirical fact that majority of the speakers who have a grasp of the concept will feel pretty confident about the truth regarding the existence of the referent of this concept. It is so simply because only few of us see ourselves as totally incapable of a rational action. Thus, it seems to be obvious that

[119] Of course, if the existential claim concerns referents of concepts that are logically impossible then it is relatively easy to deny conclusively the existence of the referents.

the world is populated by rational actions and people who take it that rational actions really exist.

Of course, it doesn`t follow from the above said that rational actions exist, but it shows what the default position is here. It shows the claim that a rational action exists to be a default position and that the burden of proof is on the denier. And once the burden of proof has been shifted where it belongs, the denier faces the full force of The Principle of Undeniability of the Rational.

(3): Free action exists.

The path from (1) and (2) to (3) is fairly straightforward. (1) tells us that rational action and free action are identical actions. (2) tells us that rational action exists. It follows that free action too exists.

Now, let me prepare the ground for the next step. I wish to show how (R) can be translated into a claim that free agents exist. We find ourselves at the point where it seems clear that free actions exist, and it might feel tempting to jump to conclusions here. If free actions exist, doesn`t it mean that free agents too exist? Unfortunately, it is not that simple. Above (p. 22), I wrote: ` An agent constitutes itself *by* acting, or ceases to be an agent entirely; `action` and `agent` are the two sides of the same conceptual coin`. I take this claim as uncontroversial. The little metaphor of `the two sides of the same conceptual coin` will help to illustrate a point here. One would expect that if something is one of two sides of the same coin, then often a property of that side will be shared by the other side too. Thus, if one side of the coin is made of silver, the other one too will be made of silver; if one of the sides is rounded, the other one too will be rounded. But, clearly, that won`t always be the case. One of the sides can have a picture of a monarch on it, while the other one can have a number there instead.

If `agents` relate to `actions` analogously (i.e. as the two sides of the same conceptual coin), we might wonder whether the fact that free actions exist implies that free agents exist or not, i.e. whether it is the case of both sides of the coin being rounded or the case of a monarch on one side and of a number on the other. In other words, we have a question to answer: does the existence of free actions entail that the agents that generate them are free?

The answer is that it depends. To see what it depends on consider the following close analogy. A *brave* action has been generated. Does it imply that the agent that has generated the brave action is herself brave (or, at least had a moment of being brave)? Not necessarily. The agent could have acted on the basis of incomplete or incorrect knowledge of the risks involved in the action. For those who have the complete – or less incomplete - picture of the risks involved, the action will clearly be a token of a brave action, the agent herself, however, might have thought that her action didn`t involve any risks to her health and safety whatsoever. In such a case, we will not want to describe the agent as brave although we will still maintain that her action was brave. The analogy shows that attributes of actions don`t automatically extend onto the associated agents. In our case it means that we cannot conclude that free agents exist just on the grounds that free actions exist. We need a separate argument here. The argument is provided below.

(4): It is an empirical fact that people, at least sometimes, act rationally (in the sense that they generate rational actions).

Above, when discussing The Principle of Undeniability of the Rational, I briefly argued that the claim that the world contains rational actions is a default position. This claim being a default position means that the burden of proof is on those who would want to dispute it. However, The Principle of Undeniability of the Rational makes any such dispute bound to fail. Disputing – to be successful – will have to be a *rational* action, which immediately defeats the attempt to undermine the default position. Now, the existence of rational *actions* entails the existence of *agents*. It is so simply because `action` and `agent` are two sides of the same conceptual coin. This is uncontroversial.

What we have here now as a default position is that a rational action – that is, acting rationally – exists and the associated agents acting (at least sometimes) rationally also exist. However, as we have seen when discussing the case of a `brave` action, the attributes describing actions cannot be automatically transferred onto the associated agents. Thus (4) shouldn`t be read as implying that the default position is that *the people* in the claim are `rational`. That wouldn`t be correct. I will, however, want to claim that, at least sometimes, it`s not only the actions that are `rational` but the associated agents too are being `rational` while generating the action. In the three steps that follow I shall argue towards this claim.

(5): An agent that generates a rational action is herself describable as `rational` if a condition *c* is met.

We know, at this point, that attributes of actions don`t automatically transfer onto the associated agents. The last sentence suggests that the transfer of an attribute – even though not automatic – is possible. The question is what exactly would make the transfer possible.[120]

[120] There is a bit of ambiguity here regarding the notion of a `transfer`. The `transfer` in question is of an epistemological kind, and not of an empirical one. The `rationality` of an agent will be instantiated, if instantiated at all, simultaneously with the `rational` action. Thus, we cannot `transfer` it if it wasn`t there from the very moment when the action was generated. But we can `transfer` it in the sense of *knowing* what needs to be the case for an agent who has generated a `rational` action to be describable as `exercising her rationality` (or simply, as `rational`) in that action. Of course, the default position for an observer of `rational` actions is that, in vast majority of cases, the agents indeed exercise their rationality in the observed (rational) actions. In other words, the default position is to always do the

Some of the bits that will go into answering this question have already been mentioned above. We were reluctant to call an agent `brave` who generated a `brave` action on the grounds that her complete ignorance of the risks may have been involved in this `brave` action of hers. Actions are `brave` when involving certain levels of risk or danger for the agent. Agents are `brave` when they are aware of those levels of risk or danger that are involved in generating `brave` actions. The condition that allows us to describe an action as `brave` - i.e. the condition that it, the action, must involve certain levels of risk and danger for the agent – must, at the same time, be a condition that in some sense constrains the intentional realm of the agent when she generates that `brave` action. The following condition is more specific about the constraints of the transfer of an attribute of an action onto the agents as it applies to the attribute of `bravery` (and perhaps some other attributes, such as: `smart`, `wise`, `cowardly`, etc.):

> Condition *c:* an attribute of an action can be transferred onto the associated agent if those constraints that an action must meet to be describable as *x* by a competent observer are, (a), constraints that the associated agent has (on a general level) a sufficient grasp of and, (b), part of her intention-formation leading to that action.

Let me unwrap the condition a bit. A competent observer – i.e. someone who has a sufficient grasp of relevant concepts and is not cognitively and perceptually impaired – will see a man running into a house on fire to save a child. She, the competent observer, has a sufficient grasp of the concept of `bravery`. She will know that what she has just witnessed is an action that can be correctly described as `brave`. She will take the action as so describable because the parameters of the action meet certain constraints of attributability of `brave` onto an action. The constraints on attributability of `bravery` onto an action are, among others, that certain levels of risk or danger for the agent must be involved in the execution of that action.

Now, condition *c* requires that if we wish to transfer the attribute of `bravery` onto the associated agent, we can do so only if that agent understands the constraints and their applicability onto actions, and these constraints are part of her intention-formation that leads to the action. Thus, the man that ran into a house to save a child will be describable as `brave` only if, (a), he understood the risks or dangers involved in running into a house on fire and, (b), he was being aware of those risks involved in running into a house on fire while forming the intention to generate that particular action.

transfer unless we have reasons not to do so. The position is a default one for the same reasons for which Davidson introduced his principle of charity.

It will come as no surprise that different attributes will be subject to different constraints of attributability. Attributability of `bravery` onto an action has essentially something to do with agent`s putting herself at risk and her being aware of the risk if `bravery` was to be transferred onto her. What about, for instance, a `dumb` action? What needs to be the case for us to be able to say that not only the action was `dumb` but the associated agent too was `dumb`? Whatever the answer here is, one thing seems to be immediately obvious: the transfer of `dumbness` from an action onto the agent will be subject to constraints that are different from those that apply to transfers of the attribute of `bravery`. While the transfer of `bravery` required the presence of a certain skill or ability on the part of the associated agent – namely, the ability to understand and appreciate the risks involved in the exercise of the intended action – the transfer of `dumbness` not only doesn`t call for the presence of any skill, it seems to call for its *absence*. And then there are cases when the attribute gets transferred fairly straightforwardly from an action to the agent. For instance, it looks like we can`t describe an action as `entertaining` without implying that the associated agent is `entertaining`. It seems that *all* that is needed to describe an agent as entertaining is simply her being the cause of an entertaining action. And she is the cause of the action already due to her being the agent in relation to that action. To complicate things yet a bit more, there are attributes that don`t transfer at all: an `unfinished` action doesn`t have a counterpart in an `unfinished` agent; similarly, there are `postponed` actions without an option of there being any `postponed` agents. And I am sure the reader would be able to come out with many other and different cases of constraints (or the absence of) concerning the transfer of an attribute from an action to the agent. There is, however, no need for us here to spend more time on these varieties. What we need instead is to look at the attribute of `rationality` in the light of what was said above and find out what (if any) constraints govern its transfer.

It seems to be fairly easy to conceive of scenarios in which `rational` actions are generated by agents who didn`t exercise any related reasoning prior to or during the `rational` action.[121] This is analogous to the case of a `brave` action. Also, `rational` is an attribute that we are totally comfortable to use when describing both actions and agents. This applies to the attribute of `brave` as well. And the analogy doesn`t stop here. Both attributes, when transferred from the associated action, seem to place a cognitive constraint onto the agent. When running into a house on fire to save a child, the agent is describable as `brave` only if she *understands* and *appreciates* the risks that her action involves. Similarly, the `rational` action in our scenario – picking up the hammer and a nail, driving the nail into the wall and

[121] Often, we don`t really exercise any robust reasoning within the time frame that immediately precedes or overlaps with the action. However, just acting in an auto-pilot mode – if that mode is anchored in some relevant reasoning that was exercised by the agent in the past – suffices to describe the agent as `being rational` or as `exercising her rationality`. For more on acting in an auto-pilot mode see section 2.5.3 in the previous chapter.

hanging the painting – is performed by an agent that is describable as `rational` in this scenario only if she *understands* a number of things here, such as: what hammers and nails are for, how paintings get attached to walls and so on. Thus, it seems safe to conclude that the attribute of `rationality` conforms to the same (the same in relevant aspects, at least) condition of transferability – condition *c* - from action to agent, as does the attribute of `bravery`.

(6): It is an empirical fact that agents often give explanations of their rational actions such that it strongly suggests that condition *c* is, at least sometimes, met.

We know, at this point, what needs to be the case for the attribute of `rationality` to be transferable from an action onto the associated agent. The constraints for the attributability are captured in condition *c*. Now the question is whether the condition *c* is, at least sometimes, met in the real world. And if it is, how could we possibly know?

Let`s start with the second question. The agent in our thought experiment seems to be engaged in an activity that looks like a good candidate for a `rational` action. The hammer and the nail are being handled in a way that is consistent with their function and causal profile. The same applies to the handling of the painting. This plus the sequence of the handling – grabbing a nail, *then* grabbing the hammer, *then* hammering the nail into the wall, *then* hanging the painting – is consistent with the interpretation of the action as a `rational` way of bringing about the desired goal, which, in this case, seems to be having the painting hanging on the wall. In other words, what we observe here is a fairly uncontroversial example of a rational action.[122]

Now, we already know that the fact that the action is rational doesn`t entail that the agent generating the action exercised her rationality in that action, i.e. that she too was rational. She might have, in fact, wanted to cook a mushroom soup and she was so confused about how to do that that she took a nail and a hammer that she saw on the table in front of her and drove the nail into the wall, and then she thought a bit more salt is needed *so* she hung the painting that happened to be on the same table on the nail. This would clearly be a case where `rationality` cannot be transferred from an action onto the agent. What is needed for a justified transfer is the agent having a desire to hang the painting on the wall, plus a sufficient *understanding* of what is involved in performing the desired action, plus this *understanding* being constitutively involved in the intention formation that precedes or runs

[122] There is an asymmetry here that should be noted. There are actions that – from the 3rd person perspective – don`t seem to be rational or might even seem outright irrational. Such actions can still turn out to be fully rational once the associated agent`s rationale becomes known. However, an action that is – from the 3rd person perspective – rational, remains so even if the associated agent hasn`t exercised her rationality in the action.

simultaneously with an execution of the action. And to find out whether all this is the case with the agent: *We just need to ask.* First, we would probably want to know why the action as a whole was performed. And we will get answers such as: `I`ve always wanted to have that painting hanging in my room` or `My daughter loves the painting and asked me to hang it in her room`. Already this kind of answer strongly indicates that the agent exercised her rationality in the action. Yet, should we perhaps want to make sure that the action was a robust exercise of rationality on other levels, we could inquire about the agent`s *understanding* of, for instance, the objects involved in the hanging of the painting. We could ask why she used the hammer to drive the nail in the wall and not, for instance, her t-shirt instead, or why she drove the nail into the wall and not into the windowpane. The answers provided will enable us to determine whether the agent did or did not exercise her rationality[123] sufficiently and on all levels that matter.[124]

Now, I don`t believe that it can be denied that people commonly give explanations of the – often rational – actions that they generate. People giving such explanations is simply a ubiquitous part of our everyday social interaction. Similarly undeniable is the fact that at least some of those explanations that relate to rational actions are *consistent* with the rationality intrinsic to those actions. By being `consistent` I mean, of course, that the explanations are such that they strongly suggest that the condition *c* has been met.

Why do I say that the existence of these explanations `suggests` that the condition *c* is met? Can`t I go for a stronger claim, perhaps something like that these explanations `show` or `prove` or `demonstrate` that *c* is met? The answer has to do with the epistemic barrier between the 3rd and 1st person realms. It is possible, at least in principle, (as shown by Searl`s Chinese-room thought experiment)[125] for an agent who lacks a capacity to *understand* things to generate an action whose behavioural profile is indistinguishable from the same action generated by an *understanding* agent. The epistemic barrier makes it impossible to establish with certainty whether the agent does or doesn`t understand her action even after she gives a rational explanation that is consistent with the rational profile of the action. The reason is that giving an explanation won`t take us beyond the behavioural realm, that is beyond the 3rd person and into the 1st person realm.

[123] Of course, there might be many cases where the agent`s exercise of rationality is only partial. In such cases it might be difficult to decide whether to describe the agent as `rational` or not. However, we don`t have to worry much about such cases as all we need for our argument to go through is that there are at least some cases where the exercise of rationality is sufficiently robust. And that seems to be very hard to deny.

[124] What is sufficient and which levels matter here will depend on one`s account of rationality and the conditions of its attributability onto agents. My argument can, I believe, accommodate any such account. Thus, there is no need to be specific here.

[125] John Searle, ‘Minds, Brains and Programs’, *Behavioral and Brain Sciences*, 3 (1980), pp.417–457.

The strength of an argument depends on the strength of its individual premises. To be able to claim that free agents exist I first need to establish that rational agents exist. And I will be able to establish that only after I show that the condition *c* is, at least sometimes, met in the real world. In (6) I manage to arrive at a claim where the evidence doesn`t `show` but only `suggests` that *c* is, at least sometimes, met. This bit about (6) will get transferred to the conclusion of the argument and weaken the strength of its claim. Instead of concluding that free agents exist I will be able to say only something like: the evidence suggests that free agents exist. Now, how much should one worry about this `weakness`?

The answer to this question is closely related to one`s theoretical attitude towards solipsism. Those who take solipsism to be a serious and respectable philosophical position will very probably be worried about the weakness. Why? The weakness that we have identified above stems from a position that is a variation of solipsism. Solipsism – the claim that the world contains only one mental subject – can be understood as a conclusion derived from combining the commitment to the existence of the epistemic barrier with the methodological imperative of Occam`s razor. In other words, we end up with solipsism once we (a) accept that the existence of any particular behaviour doesn`t entail the existence of the mental and (b) apply Occam`s razor refusing to stipulate the existence of the mental (apart from, of course, the existence of the Occam`s razor applying subject). Now, solipsism is a claim about the existence of *the mental* outside of the subject. The variation of solipsism that the weakness stems from concerns the target of the claim. In the variation it is not about the mental as a whole but only about one of its aspects, namely *understanding*. The background logic remains the same: no particular behaviour entails that the agent *understands*, and Occam`s razor bars a stipulation of *understanding*.

Solipsism is – for various reason that I won`t discuss here – a deeply unpopular position among philosophers. My presenting the weakness as grounded in a variation of solipsism, i.e. as something that is a weakness only for a (disguised) solipsist exposes the theoretical costs of targeting this particular weakness, costs that few philosophers are willing to cover. Thus, although I shall carry on saying that the evidence merely `suggests` that *c* is, at least sometimes, met, I will stay confident that not many will perceive this as a weakness in the argument.

(7): Rational agents exist.

(7) follows straightforwardly from (4), (5) and (6). No further discussion is needed.

(8): For an agent, being rational suffices to be describable as `free`; that is: if an agent is (being) rational, she is free.

At this point we understand, among others, the following two things: the logic behind describing *rational* actions as *free* actions and the constraints on the transferability of `rationality` from an action onto the associated agent. Combining the two things will give us (8). The little argument bellow shows how:

i. Agents are rational if, (a), they generate a rational action, (b), in relation to that action, *c* is met.
ii. `Freedom` is transferable from an action onto the agent.
iii. Agents are free if, (a), they generate a rational action, (b), in relation to that action, *c* is met.
iv. The identity of conditions (a) and (b) in (i) and (iii) implies that: if an agent is (being) rational, she is free.

Let me unpack the argument. (i) is a version of the claim in step (5). All that has been said in support of (5) applies to (i). (ii) should be uncontroversial. I can`t think of any semantic barriers preventing us from talking about `free` agents. In our everyday interaction with other people we both presuppose and attribute `freedom`. And if `freedom` is transferable we might wish to know what conditions constrain the transfer. (iii) gives us the conditions, and they are identical to those that govern the transfer of `rationality`. The fact of the identity of the conditions implies (8), i.e. the claim that if an agent is (being) rational, she is free; which is what we conclude in (iv).

Now, of course, the key step is step (iii). The step contains an implicit claim and two explicit claims. The implicit claim is that agency – with all the attributable descriptions – is exercised or happens to exist only in action. Any talk of a `free` agent requires having an agent in the first place, and agents are constituted by their actions. Thus, we start from an action. Telling us what action it must be is the business of the first explicit claim, which is the claim formulated as condition (a). (a) prescribes that the action must be rational. Above, we have established that rational actions are free actions so (a) could, alternatively, prescribe that the action must be free. I have opted to formulate (a) in terms of `rationality` as it fits my intention to make clear that transfers of `freedom` from agency onto the associated agents are governed by exactly the same constraints that govern the transfers of `rationality`. The second explicit claim – condition (b) – demands that the attribute of an action that we wish to transfer is transferrable onto the associated agent only if *c* is met. A reader with a good memory will, at this point, be able to see why the same condition – condition *c* – that governs the transfer of `rationality` governs also the transfer of `freedom`. The main argument behind my account of freedom in agency starts with an investigation regarding the structure of agency that needs to be in place for `freedom` to be attributable. In Chapter 2 – where I conclude that freedom is attributable only to agents who exercise practical reasoning – I argue that an agent exercises her freedom only if that exercise[126] has an appropriate rational structure. I don`t elaborate on what exactly needs to be the case for that exercise to have a *rational* structure. I am not committed to a particular theory of rationality. My account can, in this respect, accommodate a number of positions. It only requires a basic consistency in sticking with one and the same theory of rationality throughout the argument and the argument-testing attributions. Thus, whatever rational structure or constraints one accepts as governing the attributions of freedom onto agents, the same rational structure or constraints will govern attributions of `rationality` onto actions, the attributions of `freedom` onto actions and – coming full circle[127] – the transference of `freedom` onto agents. Condition *c* just restates the rational structure or constraints in the context of *transference* of `freedom` onto the associated agents.

(9): Free agents exist.
The conclusion is straightforward: if (7), rational agents exist, and (8), rational agents are free agents, then (9), free agents exist.

3.5 Conclusion

In this chapter, I have presented a thought experiment designed to show that the conclusion arrived at in the previous chapter – that is, the conclusion that freedom is attributable only to agents who exercise (practical) reasoning – is interpretable as a *sufficient* (and not just a necessary) condition of attributability of freedom to agents. The availability of such an interpretation allows us to say that a person is a free agent if and only if she exercises (practical) reasoning. And because nothing else is required, nothing prevents us from, at the same time, embracing the truth of causal determinism. This, I claim, is a robust compatibilist solution to the problem of free will. I see the proposed solution as potentially satisfactory for a specialist in the problem of free will. A non-philosopher might, however, feel cheated as the proposed solution seems to be confined to a purely theoretical realm of conceptual relations. Thus, to a non-philosopher, I have offered an expanded interpretation concluding that: free agents exist. This is a kind of conclusion that, I believe, is what a non-philosopher would be looking for as a satisfactory outcome of attacking the problem of free will.

[126] `The exercise of freedom` here refers not to an (extra-mental) action but to the act of making a choice as part of the intention-formation that precedes that action.

[127] The circle is not vicious because (C) is not an assumption but a claim that is a result of a conceptual analysis.

Chapter 4: Holding responsible and the free will assumption

4.1 Introduction

There is a problem regarding the justification of our common practice of holding others morally responsible (PHMR). As far as I am aware, all philosophers working on the problem of free will understand the concept of free will and the related problem as something that is to be approached and resolved within the conceptual and theoretical constraints of (PHMR). In Chapter 1, Freedom and morality: severing the connection, I argued that this approach is deeply misleading. The concept of free will and the related problem is independent of (PHMR) and should be approached as such. Severing the connection between the discussion of free will and (PHMR) has allowed me to develop (in Chapter 3) a novel compatibilist account of freedom in agency. The account shows in what sense it is plausible to attribute freedom to reasoning agents despite them being fully causally determined.

Now, the account of freedom in agency that I have arrived at is such that it will be of no help regarding (PHMR), i.e. it won`t ground justifications of holding others moral responsible. Why won`t it do the grounding? Philosophers theorizing about free will within the framework of (PHMR) - that is, as far as I know, all of them who theorize about it – will deem an account of free will successful only if it generates intuitively correct results when employed in moral stories. The results will be taken as correct if they accord with our moral intuitions as elicited by the moral stories. Most philosophers` attempts to formulate their accounts of free will in terms of *control* or *sourcehood* reveal what is felt as intuitively constraining our moral intuitions when tested in relevant moral scenarios: typically, we seem to be reluctant to hold a person morally responsible for her conduct if the person lacks an intuitively satisfactory kind of *control* over her actions and/or is not – in some, again, intuitively satisfactory way – the ultimate *source*[128] of her actions.

The reader will remember that in Chapters 2 and 3 my approach to the discussion of the problem of freedom in agency was separated from any theoretical considerations related to (PHMR). This means that the development of my account hasn`t been constrained by those moral intuitions that require an account of free will to be built around the concept of *control* and/or *sourcehood*. And once the discussion of freedom in agency has been disconnected from (PHMR), there was no reason whatsoever to voluntarily introduce the constraints. No wonder then that once the account gets tested in relevant moral scenarios, we get wrong

[128] Kevin Timpe argues that the kind of control required for morally relevant freedom is, ultimately, analysable in terms of sourcehood. If he is right – and I believe he is – then our moral intuitions as elicited by the thought experiments will ultimately scan the thought experiment scenario to determine whether the sourcehood is constituted and if it is, to what extent it has been corrupted by circumstances introduced by the scenario. For details, see Kevin Timpe, *Free Will: Sourcehood and Its Alternatives* (Continuum, 2008).

results. Attributions of freedom onto agents will be justified – in the context of my account - in various cases where holding another responsible would be felt as incorrect, that is, in cases where our moral intuition detects lack of satisfactory *control* and/or *sourcehood* on the part of the agent.

Disconnecting the discussion of the problem of free will from the theoretical commitments related to (PHMR), which was done in Chapter 1, could have been taken to be just a sort of a dialectical step intended to allow a new approach to the solution of the problem. Thus, it might have been felt that, perhaps, later it would be possible to reconnect the account with the issue of (PHMR). Perhaps, the account could be developed such that it would do the required theoretical work within the framework of (PHMR). Surely, that`s what we all expect from a successful account of freedom in agency. And should it turn out to be impossible to develop the account in this way, it would, I concede, be taken by many, or even most, philosophers to be a major weakness of the account.

A natural response to this perceived weakness would be to try to find out whether it could be developed further into an account that lacks the weakness, i.e. an account which would be able to ground the ascriptions of moral responsibility. There are, however, serious reasons to doubt that an account of free will robust and plausible enough to justify our common practice of ascribing moral responsibility can be defended.[129] Thus, instead of trying to develop the account to fit the needs of a moral theorist, I shall problematise the ethical rationale behind our common practice of holding people morally responsible.[130]

I will analyse the concept of *holding morally responsible* and identify some important ambiguities. The analysis and the resulting identification of ambiguities will allow me to argue that all the *right* holdings of moral responsibility can be performed without the need for the free will assumption.[131] It will turn out that only the *wrong* holdings require the free will assumption. This distinction regarding attributions of moral responsibility will remove the apparent weakness of my account discussed above. I will proceed in the following way. I will analyse the concept of holding morally responsible. With each identified reading, I will examine whether or not it calls for the free will assumption. Before that, however, I will need

[129] The main problem is, very briefly, this. The ascriptions of moral responsibility (of the kind that call for the free will assumption) can be, ultimately, justified only if the free will available to agents involves the power of *causa sui*. Successful explanations are formulated in terms of constitutive reasons and/or causes. An explanation that refers to the mysterious power of *causa sui* violates this requirement and will, therefore, not be considered successful. Put simply, explaining something in terms of *causa sui* amounts, ultimately, to not explaining it at all.

[130] The assumption here is that if the practice of holding others morally responsible is, on separate grounds, exposed as seriously problematic then any theoretical requirements stemming from it will lose their appeal.

[131] The free will assumption referred to here is the one in which *freedom* is understood to do essentially with *control* over actions or with being the ultimate *sourcehood* of them.

to discuss an issue that has been, until now, assumed unproblematic. At this point, the issue will be merely touched upon and later, when more will turn on it, it will be discussed in more detail.

4.2 A clarification regarding the notion of justification

The issue is this. To be able to examine whether the practice that a particular reading refers to *calls* for the free will assumption, we need to have some understanding of what constitutes such a *call*. A practice referred to by a reading of the concept of holding morally responsible can, of course, *call* for anything only in a metaphorical sense. The *call* is best to be understood here in terms of *justification*. Thus, the question whether a particular practice *calls* for the free will assumption can be rendered as a question about whether a particular practice requires a justification of a sort that depends on the free will assumption.

Now, the whole discussion of the problem of free will is motivated by the shared understanding that at least some varieties of our practice of holding others morally responsible stand in need of such a kind of justification, i.e. one that depends on the free will assumption. That much is taken, by the majority of philosophers, to be uncontroversial. Interestingly, the very notion of a `practice standing in need of a justification that depends on the free will assumption` seems to be taken as transparent and unproblematic. The literature on the problem of free will characterizes the problem as a pressing one on the grounds that what is at stake is the justification of our common practice of holding people responsible: if we cannot assume that people are free agents, we cannot justify our practice of holding them responsible as agents and for their conduct. At exactly this point it is left unexplained what the procedure of *justification* itself involves. It seems to be taken as intuitively clear.

I suspect, however, that the issue of *justification* here is much less unproblematic and transparent than is commonly assumed. There are at least two considerations that support my suspicion. First, there are different kinds of justification. As there is no need to elaborate here, I shall mention, as an example, just two most obvious kinds: an epistemic justification and a moral justification. The employment of each requires different logic. My point here is that if there are different kinds of justification, governed by different logic, then we need to be clear about the inner workings of the governing logic as it applies to each different kind. Second, as mentioned above, the notion of `holding morally responsible` has several readings. Each of these readings refers to a different practice. It is reasonable to expect that different practices might differ also with respect to the exact nature of the justification requirement that applies to each of them. It might, and it will, turn out that not all of the practices picked up by the notion of `holding morally responsible` stand in need of the justification grounded in the free will assumption (although they might stand in need of some other kind of justification).

Now, of course, to be able to recognize which practice stands in such a need and which doesn`t, we have to be clear about what must be the case for something to achieve the status of being *morally* justified. However, what I mean here by being `clear about what must be the case` is not being clear about any particular details (that are different in different cases) whose presence or absence will determine whether that something (a conduct) is morally justifiable. What I mean by the phrase is, again, the clarity about the underlying logic that is employed to process the particular details and generate a verdict. Only when we are clear about the underlying logic shall we find ourselves in a position to distinguish those holdings that are morally justifiable only under the free will assumption from those that are in no need for the assumption. And I don`t think we have this clarity about the underlying logic.

So, what is that underlying logic that governs our inquiry into whether a particular exercise of holding another responsible is morally justified? I will claim and (later) argue that what we are intuitively looking for when evaluating the moral justifiability of a conduct is whether *fairness* (in some morally significant sense) has been preserved. At this point, however, I will just rely on my reader`s intuitive grasp of the notion of `fairness` and ask – about any particular kind of practice of holding responsible – whether or not it stands in need of moral justification, i.e. whether or not it is about fairness. If a particular kind of practice turns out to be about fairness, I shall assume it calls for the free will assumption.

I concede that assuming so might be seen as somewhat unwarranted here. It might be seen as such to the extent to which it turns on a cluster of implicit assumptions, such as: holding morally responsible relates primarily to wrongdoings; holding morally responsible essentially involves the imposition of pain on a wrongdoer; imposing pain on a wrongdoer stands in need of moral justification; imposing pain for a wrongdoing is morally justifiable – that is, fair – only if the wrongdoer could have done otherwise, that is, only if he has free will. I believe the assumptions do not need to be discussed individually and explicitly here. It suffices if the reader agrees – as she will, I believe - with a strong moral intuition that it is *unfair* to punish a wrongdoer who lacks free will.

4.3 Holding morally responsible

Talk of `holding morally responsible` is ambiguous on a few levels. Let me first deal with two issues of less importance which, nevertheless, could, if left unidentified, obscure those aspects of the notion of `holding morally responsible` that are the actual target of my analysis here.

4.3.1 *Holding* vs *attributing*

First, I wish to explicitly distinguish *holding* a person morally responsible (for a conduct) from *attributing* moral responsibility (for a conduct) to a person. *Holding* a person morally responsible is – in a sense explicated below – a richer act than just *attributing* moral responsibility to her. The former implies the latter but not vice versa. *Attributions* of moral responsibility are possible without being accompanied by holdings (an unexpressed identification of a culprit behind a morally reprehensible action, for instance) and often it would be unintuitive to perform holdings at all (in case of non-agents responsible for something) while attributing. *Attributions* of responsibility don`t call for the free will assumption for reasons elaborated on below when Strawson`s concepts of *reactive attitudes* and *objective stance* are discussed. Here it suffices to notice that with attributions of moral responsibility for a conduct we find ourselves on the territory of epistemic justifications and not on that of moral justifications. And while the moral justifications have, as mentioned above, something to do essentially with *fairness*, the epistemic ones are essentially related to truth. In other words, with attributions we don`t ask whether they are *fair* but whether they are *correct*.

4.3.2 Moral vs non-moral responsibility

Second, there are holdings responsible which are not *morally* loaded. A person can be held responsible in the sense of being placed `in charge` of something, which is a kind of holding responsible that is not morally loaded (at least not directly). We commonly say things like: `The person responsible for our computer system is John Smith`; or ask: `Who is responsible for the promotion and advertising in this firm?`. In the two sentences above, being `responsible` means being `in charge` (with all related duties and obligations). Here too, the usage of `responsibility` carries no moral significance, or, at least, no *direct* moral significance.

The qualification `direct` is meant to pre-empt a possible objection: Surely, John Smith, who is responsible for the computer system can be held responsible, i.e. *blamed* or *praised* for malfunctioning or smooth performance respectively of the computer system. John Smith is – in virtue of him being *responsible* for the computer system – open to attitudes of *blaming* and *praising,* which indicates that this kind of responsibility is *morally* significant. To answer the objection, we need to distinguish between duties and obligations that are about *moral* rights and wrongs and those that aren`t. Holding a person *morally* responsible is, among other things, a *binding* act, i.e. an act that imposes a duty or obligation to behave in a *morally* right way. Holding a person responsible for a computer system is also a *binding* act; here too we impose a duty or an obligation onto a person. The difference here is that in the case of holding a person *morally* responsible the duty and obligation we impose onto her has to do essentially with *moral* rights and wrongs while in the case of holding one responsible for a computer system the duties and obligations have to do essentially with the *performance* of the system. Clearly, the performance of a computer system is something that is subject to essentially *non-moral* criteria. It is about being fast, reliable, accessible, secure, virus-free, etc. Responsibility for a computer system is related to outcomes in a *non-moral* domain, while *moral* responsibility is related to outcomes in the *moral* domain. Thus, a person responsible for a computer system can be praised as someone excelling at the fulfilment of her responsibilities on the grounds of the system running flawlessly. All this despite the fact that the computer system itself is used by the military of a tyrannical state to help wage a genocidal war onto a peaceful country, and our computer system manager is fully aware of this fact.

Of course, the fact of being responsible for a computer system doesn`t mean that one cannot be held *morally* responsible in connection with that computer system. To the extent to which the computer system affects other people`s wellbeing, it can be *used* to bring about morally significant effects. And for those morally significant effects the manager of the system is to be held *morally* responsible. The above-mentioned case of the computer system being used for waging an illegal war with full awareness of the manager of that system is a good example. And a war crime tribunal will not hesitate to hold the manager morally responsible and punish accordingly. A less extreme example would be one of a malicious manager of the computer system who dislikes his colleagues and intentionally bugs the system to annoy them. Here, in this less extreme case, two things might get conflated: the manager will be held *non-morally* responsible for the poor performance of the system but also *morally* responsible for intentional malevolence towards others. Yet, the two kinds of responsibility shouldn`t be conflated. Holding someone responsible for helping to wage an illegal war or for intentional malevolence towards her colleagues through bugging their computers is clearly different from holding someone responsible in the sense of placing them `in charge` of something. The former belongs to the class of *reactive attitudes* and shall be discussed in more detail below.

Now, does *holding responsible* in the latter sense call for the free will assumption? I don`t think it does. Holding responsible in the sense of placing a person in charge of something (with all the duties and obligations that come with it) is placing them in a certain causal relation to that something. It amounts to imposing on them the role of acting as a cause of a (desirable) performance of that something. We impose these causal roles on people for practical purposes. For instance, the level of efficiency and reliability of something`s performance will depend on the existence and performance of its cause(s), i.e. on the existence of the manager (in charge) and his performance. Systems without managers or with poor managers perform much worse than systems with good managers. In case of poor performance or malfunctioning of the system the existence and identification of the person in charge will often allow for fast and effective fixing of the problem. The person in charge is, typically, well positioned to understand and access the system she is in charge of, which makes her the person to call upon if the system needs to be fixed, its performance improved or an explanation related to it is required.

It might be conceded that holding responsible in the sense of placing in charge doesn`t stand in need of the justification grounded in the free will assumption. At this point, we might want to know whether this kind of holding responsible stands in need of *any* justification at all. I think it does. The kind of justification called for here will be some version of a simple epistemic justification. Consider the following scenario:

You are a manager of a big company. One of your numerous responsibilities is to choose and place people in charge of various departments. There are many departments and, consequently, many people to choose and place in charge. No wonder that you are sometimes unsure who is in charge of which department. This time your memory seems to have failed you with respect to who the person responsible for the computer system in your company is. You ask yourself: Is John Smith the person I am holding responsible for the (performance of the) computer system?

I think the context of this little scenario makes it clear that what the question really asks is: is John Smith the person (I placed) *in charge* of the computer system? Now, the answer given to the question will depend on whether it is *true* or not that John Smith has been placed *in charge* of the computer system. The answer will be affirmative in case the identity of the person in charge is John Smith, negative if it is someone else. Nothing else is needed, that is, it doesn`t need to be established whether any holding of John Smith to moral oughts has occured. The absence of a holding to moral oughts implies that we don`t enter the territory of *fairness* with this kind of holding responsible.[132]

4.3.3 Reactive attitudes vs objective attitudes

[132] I admit that the distinction between holding responsible in the sense of placing in charge and in the sense of holding to moral oughts is a bit slippery. The reason is that the two senses rarely, if at all, come apart in real life situations. Placing someone in charge opens that someone to being held to moral oughts. And it is, perhaps, inevitable that sooner or later the holding to moral oughts will – in some form, weaker or stronger – take place. The whole issue gets complicated by the fact that placing someone in charge involves binding or holding of sorts. However, this binding or holding is not to *moral* oughts but just to oughts related to the performance of whatever it is that the person has been placed in charge of. This binding or holding to non-moral oughts together with the fact that the two senses rarely come apart in real life situation is what somewhat obscures the distinction.

We have distinguished *holdings* from *attributions*, and *morally loaded* holdings from those holdings that are not thus loaded. We are interested in morally loaded holdings. Here another important distinction needs to be made. We can hold someone responsible for her *conduct* or hold her responsible in the sense of *regarding* her as a responsible *agent*. The distinction lies at the heart of the argument that Strawson presents in his seminal article `Freedom and Resentment`.[133]

Holding someone responsible in the sense of regarding her as a responsible agent is a stance enabling us to engage with her as a capacitated agent. It is a broad psychological attitude towards another which shapes our expectations and presumptive interpretations, and leaves us predisposed to certain reactions and interactions. Strawson calls the reactions and interactions that we are predisposed to when acting from within this stance `reactive attitudes`. This stance is a default attitude we have towards others in ordinary personal relationships.[134] It is a default attitude in a deep psychological sense. This attitude allows us to get immersed in rich interpersonal relationship by regarding others as capable of the same immersion. I will follow Coleen Macnamara in calling this kind of holding responsible the `participant stance`.[135]

The participant stance is to be distinguished from the objective stance.[136] The objective stance is a stance that we take towards people who are deficient in some relevant agential way, or generally towards non-human animals and inanimate objects. We can also take this stance towards people in institutional or theoretical contexts where they are treated as objects or abstractions. The expectations, interpretations, reactions and interactions that the objective stance predisposes us towards are importantly different from those associated with the participant stance. The reactive attitudes – that is, the attitudes associated with the participant stance – are *essentially* emotional. They are, for instance, `resentment, gratitude, forgiveness, anger, and the sort of love which two adults can sometimes be said to feel reciprocally for each other`.[137] The participant stance assumes a robust agency on the part of the target of the stance. The objective attitudes – the ones associated with the objective stance – are not emotional in an *essential* way even though they may be `emotionally toned in many ways`.[138] The objective attitude `may include repulsion or fear, it may include pity or even love, though

[133] Peter F. Strawson, `Freedom and Resentment`, in John M. Fisher and Mark Ravizza (eds), *Perspectives on moral responsibility* (Ithaca: Cornell University Press, 1993), pp.45-66.

[134] Ibid., p.55.

[135] See Coleen Macnamara, `Holding others responsible`, *Philosophical Studies*, 152, no. 1 (2011), pp.81-102 (p.82).

[136] Strawson doesn`t employ the notion of `stance` to draw the distinction. Instead he talks about the `attitude` (of participation). The attitude of participation can be, in some cases, suspended and replaced by the objective attitude.

[137] Strawson, `Freedom and Resentment`, p.52

[138] Ibid.

not all kinds of love`.[139] Say, you have found a baby seagull that had fallen from the roof. Apparently, it has been abandoned by its parents and can`t fly yet. Your stance towards the baby seagull will be objective and yet `emotionally toned`. You might suffer from some kind of phobia regarding birds, therefore, you might feel *repulsion*. You might think the bird is ill and *fear* that it could infect you if it touches you. You might be struck by the helplessness of the poor creature and feel *pity* for it. Consequently, you might take it home, care for it and, by the course of time, come to *love* it as your pet. However, your stance towards the seagull will lack certain dimensions (typical for the participant stance) because of the absence of a robust agency on the part of the seagull.

For our purposes, the key difference between the reactive attitudes and the objective attitudes is that some of the former are *morally* loaded while none of the latter ever are. Accepting another`s promise, offering her confidences or falling in love with her are examples of robust reactive attitudes that are *not* morally loaded, while feeling resentment, indignation, gratitude or forgiveness for someone are examples of reactive attitudes that are morally loaded. The objective attitudes are never morally loaded because through them we attribute a responsibility that is of an explanatory and/or causal kind, or is a mere *grading*.[140] Thus, the objective stance and the associated objective attitudes shall be (together with holdings responsible in the sense of placing someone *in charge* of something) put aside as *holdings* responsible that are, for our purposes here, irrelevant because they are not morally loaded. At this point, the reader will have already understood that not being morally loaded implies having nothing to do with *fairness*, thus not calling for the free will assumption.

4.4 Reactive attitudes: entering the moral territory of fairness

What we are then left with – when attempting to disambiguate the notion of `holding morally responsible` - is the participant stance and the related reactive attitudes. Our actual target here are, ultimately, the reactive attitudes. The participant stance, being a *stance*, is merely a *dispositional* state. It is structurally rich but essentially *inert*. From without the participant stance we *take* or *regard* others as responsible but this *taking* and/or *regarding* is behaviourally *unreactive*. Adopting the participant stance, we don`t really *do* anything to the other; we don`t *reach* towards the other. And it is this absence of *reaching* towards the other that is incompatible with *holding* the other to any moral oughts.[141] Also, and perhaps more

[139] Ibid.

[140] Attributing responsibility as *grading* is a distinction made by J. J. C. Smart in his "Free Will, Praise and Blame", *Mind*, 70, no. 279 (1961), pp.291–306. *Grading* is an attitude towards another involving an attribution of responsibility which assesses another against her own evaluative standpoint, her practical identity, and what they `stand for`. It doesn`t involve attributions of *moral* responsibility.

[141] This point will be elaborated on below when the distinction between holding responsible as *deep moral appraisal* and holding responsible as *holding accountable* is introduced and discussed.

importantly, the essential *inertness* of the participant stance means that *fairness* – should it be raised as an issue at all – stays unaffected. And if this is the case then we can safely conclude that the practice of adopting the participant stance doesn`t call for a theoretical justification grounded in the free will assumption.

Let me briefly address a possible objection to this conclusion: The participant stance is a stance of regarding another as a *responsible agent*. And it seems reasonable to insist that regarding someone as a *responsible agent* entails regarding her as *an agent that possesses free will*. That`s what majority of philosopher working on the problem of free will think anyway. Thus, it looks like the participant stance does call for the free will assumption after all. And if that is the case then the reactive attitudes will stand in need of this assumption too, as they are intelligible only against the background of the participant stance. Two things can be put forward to counter the objection.

First, the fact (if it is agreed to be a fact) that regarding someone as a *responsible agent* entails regarding her as *an agent that possesses free will* doesn`t call for a theoretical justification grounded in the free will assumption exactly because it already is, as a matter of fact, *entailed* in the participant stance. The participant stance doesn`t call for the free will assumption not because it doesn`t need the assumption *but* because the assumption is already there. The status of the assumption is pre-theoretical in the sense that its being there is not a result of theorizing but of some other, non-theoretical processes.

Second, there is an influential and coherent account of *responsible agency* that is independent of the free will assumption. The account has been formulated by Harry Frankfurt in his `Alternate possibilities and moral responsibility`.[142] Frankfurt`s argument turns on a thought experiment designed to show that the moral responsibility of an agent is not diminished by a lack of alternatives. If we define free will in terms of the availability of alternatives, as most philosophers do, we can, countering the objection, appeal to Frankfurt and claim that others are to be treated as responsible agents even if they possess no free will.

Thus, a Strawsonian has a range of plausible theoretical options to defend her claim that the participant stance (and the related reactive attitudes) do not call for a theoretical justification grounded in the free will assumption.

4.4.1 Reactive attitudes and the free will assumption

It is only with the reactive attitudes that we *do* something to the others, that we reach to them. And it is only with the reactive attitudes (not with all of them though) that we hold others responsible in the sense of holding them to moral oughts. Have we then finally identified the

[142] Harry G. Frankfurt, `Alternate possibilities and moral responsibility`, *The Journal of Philosophy*, 66, no. 23 (1969), pp.829-839.

kind of holding responsible that is sensitive to the requirement of *fairness* and that, as such, calls for the free will assumption?

Strawson, who introduced the concept of reactive attitudes in his `Freedom and Resentment`, famously argues that it doesn`t. Let me briefly recap Strawson`s position. Strawson attacks the widely shared assumption that holding persons responsible rests upon a theoretical judgement of their being responsible. Strawson takes as uncontroversial that being responsible presupposes having free will. Thus, his attack is to be understood as directed against the assumption that holding persons responsible rests upon a theoretical judgement of their having free will (or being free agents in some sufficiently robust sense). Strawson claims that we find the problem of free will to be a pressing issue only because we have over-intellectualized the issue of moral responsibility. The over-intellectualization consists in assuming that the rationality of our practice of holding another responsible depends upon a judgement that the person held responsible has satisfied a set of objective requirements, where having free will is, typically, taken to be one of the objective requirements. Strawson argues, however, that our practice of holding others responsible doesn`t depend on any such theoretical judgement. Our practice of holding others responsible, as he puts it, `neither calls for nor permits, an external `rational` justification`.[143]

The argument behind this claim is, roughly, that the practice of holding others responsible is essential to and constitutive of our social interactions. It cannot be given up without giving up what we are as social beings. Our psychological and social constitution is such that it is impossible for us to divest ourselves of this practice. Moreover, even if it was possible to suspend our reactive attitudes (where some of these reactive attitudes are forms of holding others morally responsible) and assume permanently the objective stance (and related attitudes), it would still remain rather doubtful that rationality could ever justify such a profound `reformatting` of human nature. Therefore, Strawson concludes, it is pointless to theorize about free will in relation to the issue of the justifiability of the practice of holding responsible. It is pointless because the practice – even if it needed any justification - is too fundamental to be susceptible to the justifying force of rationality.

4.5 Reactive attitudes and the free will assumption: contra Strawson

If Strawson is right we could and should dismiss all reactive attitudes - and with them the remaining kind of holding morally responsible – as standing in no need of a justification grounded in the free will assumption.[144] I believe, however, that Strawson is only partly right.

[143] Strawson, `Freedom and Resentment`, p.62.

[144] This is, strictly speaking, not what Strawson is saying. Strawson could be very plausibly taken as saying the exact opposite: The free will assumption cannot be given up exactly because it comes part and parcel with the participant stance and reactive attitudes, which is something that we are unable to

More specifically, I want to claim that some reactive attitudes – and with them a certain kind of holding morally responsible, (a), can be given up, (b), require justification that is grounded in, among others, the free will assumption. Arguing towards the claim, I will rely heavily on a distinction made by Coleen Macnamara in her `Holding others responsible`.[145]

4.5.1 Holding responsible as deep moral appraisal (HRDMA) vs holding responsible as holding accountable (HRHA)

Macnamara notices that contemporary theorizing about the concept of holding responsible is plagued by quite a few unresolved questions. Does holding responsible involve *both* sanctioning behaviour and praise, or sanctioning behaviour only? If it is both, then why does nearly all the discussion revolve around sanctioning behaviour, and very rarely around praise, as examples of holding responsible? What explains this conspicuous asymmetry? What is the relation between holding a person responsible as an *agent* and holding her responsible for her *actions*? What is the relation between holding someone responsible and the participant stance? Are the two equivalent, different, one part of the other? What is the status of *unexpressed* reactive attitudes and normative expectations? Do they both count as holdings responsible?

Macnamara argues that the questions keep popping up because we lack certain crucial distinctions. Once the distinctions are made, the questions can be answered and put to rest. The distinction that I wish to make use of here concerns holdings morally responsible for a particular piece of conduct. That is, the distinction concerns the morally loaded reactive attitudes.

Macnamara distinguishes between holding responsible as *deep moral appraisal* (HRDMA) and holding responsible as holding *accountable* (HRHA). Both faces of holding responsible are important in their own right. They often come merged in a single act of holding another responsible, yet they should not be conflated.

(HRDMA) are essentially psychological reactions to morally significant conduct of others. `We *feel* a reaction to someone`s action; we *express* a reaction`.[146] I despise my neighbour for cheating on his wife: `You`ve got kids, man. Stop thinking with your private member!`; I disapprove of my friend for driving a big SUV instead of a small and more environmentally friendly car: `You don`t seem to have heard about global warming, mate,

suspend permanently. The `cannot` in `cannot be given up` is not a normative one but a modal one: it is a `cannot-because-impossible` kind of `cannot`. However, once it has been established that something cannot be altered it makes any attempts to justify alternations blatantly pointless. That which cannot be altered stands in no need of justification regarding its conceivable alternations.

[145] Coleen Macnamara, `Holding others responsible`, *Philosophical Studies*, 152, no. 1 (2011), pp.81-102.

[146] Macnamara, `Holding others responsible`, p.89.

do you?`; I praise my daughter for giving up her seat to an elderly lady: `That was a very kind thing to do`.

However, we often do, or wish to do, more in the act of holding another responsible than just emotionally react to another`s morally significant conduct. We hold others responsible in the sense of holding them *accountable* (HRHA). Macnamara gives an example of a worker who is having a difficult time at work.[147] Imagine the worker is your sister. The boss distributes the workload unfairly and your sister has to do more than others. Moreover, she is the one who always gets blamed by her boss when things go wrong and never gets any credit when success comes her way. She is afraid that things might get worse if she openly expresses her resentment towards her boss, so she keeps it bottled up inside. This unfair treatment of your sister drives you crazy. You are trying to convince her that she needs to stand up for herself: `Don`t be so damn passive! You can`t tolerate this any longer! Stand up for yourself! *Hold him responsible!*`.

Macnamara argues that your remark would make no sense if moral appraisal was all we do when holding others responsible. `Why would you urge your sister to do something she is already doing?`, Macnamara asks.[148] Your sister already *resents* her boss which is an act of moral appraisal. And it is an act of moral appraisal even if it stays unexpressed. However, it all starts making sense once we understand the `holding responsible` here as referring to *accountability*. Your sister is being urged to hold her boss *accountable*.

But what exactly is it to hold someone accountable? Macnamara argues that the best way to grasp this notion of holding responsible is via `the metaphor of *holding* someone to the oughts that bind them`, where the notion of *holding* `is best understood on the model of enforcement`.[149] It works like this. When holding someone *accountable* what we do is `perform a communicative act with a distinct internal aim`.[150] The internal aim is to:

> [I]nduce what we might call first-personal practical uptake of the ought-violation in the one we`re holding accountable - to get the wrongdoer to acknowledge her wrongdoing, feel remorse, apologize, make amends, and commit to doing right in the future.[151]

To achieve this internal aim, we impose *burdens* - `the pain of punishment, the sting of reproof`.[152] Macnamara illustrates with the following example:

> [I]magine that you are an avid environmentalist taking a walk with your friend. As you stroll along, your friend takes out a candy bar, unwraps it, and blithely throws the wrapper on the ground. Appalled, you lay into him; you reprove him. Your reproof is, I want to argue, a distinctive kind of act. It is a communicative act that aims at inducing in your friend first-personal practical uptake of his wrongdoing. The point of your speech act is to get your friend

[147] Ibid., pp.89-90.
[148] Ibid., p. 90.
[149] Ibid.
[150] Ibid.
[151] Ibid.
[152] Ibid.

> to recognize that he has done wrong, to feel remorse, to apologize and make amends, and to commit to not littering in the future. And it does this, not by merely pointing to or highlighting the ought-violation—displaying it or calling his attention to it. It does so, instead, by imposing burdens on your friend—in this case, the sting of the rebuke.[153]

Now, I would like to believe that, at this point, the distinction between (HRDMA) and (HRHA) is sufficiently clear. (HRDMA) is essentially an emotional reaction to a morally significant conduct. (HRDMA) can come both expressed and unexpressed because appraisals are accomplished even when unexpressed. (HRDMA) applies not only to blameworthy conduct but to praiseworthy conduct too because appraisals commonly target morally *right* actions, not only the *wrong* ones. (HRHA) goes beyond mere appraisal. It aims to bring about a change in the one held responsible, namely, `first-personal practical uptake of his wrongdoing`. The uptake is *enforced* through imposing *burdens*. (HRHA) must be *expressed* in some way to achieve a successful uptake, and it applies to blameworthy conduct only. (HRHA) applies to blameworthy conduct only because unlike the praiseworthy conduct the blameworthy conduct reveals a moral deficit on the part of the agent. And aiming to bring about a practical uptake presupposes such a deficit has been revealed. Thus, we can clearly see at this point how distinguishing (HRDMA) – that is, holding responsible as a deep moral *appraisal* – from (HRHA) – that is, holding responsible as holding *accountable* – helps answer some of the questions that Macnamara identifies as crucial if progress is to be made. Does holding responsible involve *both* sanctioning behaviour and praise, or sanctioning behaviour only? It involves both when we engage in (HRDMA), and sanctioning behaviour only if we engage in (HRHA). What is the status of *unexpressed* reactive attitudes and normative expectations? Do they both count as holdings responsible? The two count as instances of holding responsible in (HRDMA) but not in (HRHA).

There is a little complication here. The complication stems from two facts.

Fact 1: (HRDMA) and (HRHA) get commonly merged in a single instance of a reactive attitude, therefore the distinction might feel somewhat artificial. Thus, when you rebuke your friend for littering the environment you not only attempt to induce the first-person practical uptake (i.e. to hold your friend *accountable*) but you express your indignation too. However, the two faces of holding responsible – (HRDMA) and (HRHA) – come apart in some ordinary situations: when praising and when keeping the appraisal unexpressed.

Fact 2 is slightly more serious: (HRHA) is, among others, a *communicative* act. A communicative act has an aim that is *internal* or *essential* to it and an aim that is *external* to it, imposed by the speaker. The two shouldn`t be confused: After two pints of strong ale, a friend of yours is getting into his car intending to drive home. `Are you mad?`, you ask. The

[153] Ibid., pp.90-91.

question `Are you mad?` is a token of a communicative act whose internal aim is to inquire. The associated external aim, however, is different. You, the speaker, are not trying to extract an information from your friend about the state of his mental health. You aim to make your friend realize that driving when drunk is wrong. In other words, it only *looks like* you are asking a question. What you really do is trying to prevent your friend from driving. How does this relate to (HRHA)? There will be instances of holding responsible that will *look like* (HRHA) while being a mere (HRDMA). You make a friend of yours privy to an embarrassing detail about your family life. The friend, however, brakes the promise of keeping that detail strictly to himself and tells an acquaintance of yours about it. Upon learning about your friend`s betrayal, you pick up your phone and call him. As soon as your friend answers your call you start shouting: `You little miserable moron! How could you do that?! You make me sick!`. Surely, this is something that sounds like a rebuke! As such it might result in the first-personal practical uptake of the wrongdoing, which is something that suggests that what we have here is an instance of (HRHA). However, (HRHA) is a communicative act and as such has both *internal* and *external* aims that we associate with communicative acts. A communicative act can be *internally* a rebuke and *externally* a mere *expression* of resentment. Expressing resentment is what you have aimed to do, and you gave it the form of rebuke. And you might have even, accidentally, accomplished the first-personal uptake, which could be taken as implying that a (HRHA) has taken place. But that`s not what you intended. In other words, you performed (HRDMA) using a communicative form normally associated with (HRHA).

We can see that these little complications can be dealt with and shouldn`t, therefore, prevent us from fully appreciating the theoretical utility of distinguishing the *appraisal* face of holding responsible (HRDMA) from the *accountability* one (HRHA).

Now, let me get back to the motivation behind my introduction of Macnamara`s distinction. Above, I promised to argue that some holdings – contrary to what Strawson claims – can, (a), be given up and, (b), stand in need of a theoretical justification grounded in the free will assumption. So how will the distinction help me to argue for (a) and (b)? It will, once suitably developed. Before doing so, let me briefly examine how (HRDMA) and (HRHA) relate, prior to the further development, to the free will assumption.

4.5.2 (HRDMA), (HRHA), and the free will assumption

Above, I conceded that it is right to say that at least some reactive attitudes cannot be given up (permanently). As such, they stand in no need of a theoretical justification that depends on the free will assumption. I take it as uncontroversial that the reactive attitudes associated with (HRDMA) are a subset of such reactive attitudes. Being essentially *emotional* responses, these reactive attitudes are constitutive of what we are as human beings and how we relate to others. They cannot be permanently suspended without causing serious disruption on the level of individual psyche and on the level of social interaction. This fact makes any concerns about their justification pointless. In this sense, the reactive attitudes associated with (HRDMA) stand in no need of a theoretical justification grounded in the free will assumption.

As for the reactive attitudes associated with (HRHA), things are, at least on the face of it, different only partially. These reactive attitudes can be given up, but it looks like they don`t – as don`t those associated with (HRDMA) - really call for any theoretical justification grounded in the free will assumption. As for the suspendability of these reactive attitudes – the ones associated with (HRHA) - the reader will recall the scenario in which their sister keeps her resentment bottled up inside never holding her boss *accountable* for his unfair treatment of her. The sister feels resentment towards her boss, that is, she holds her boss responsible in the sense of (HRDMA). She, however, refrains from holding her boss responsible in the sense of (HRHA), and it is far from inconceivable that she will be able to do so for a long time or even for the rest of her life.

A little complication should be noticed here. We might feel that keeping the resentment towards one`s boss bottled up inside is not really a sustainable situation. One doesn`t need to have a degree in psychology to know that supressing one`s emotions is not only mentally unhealthy but unsustainable in the long run. That`s a fact about our psychological constitution. Isn`t this exactly what Strawson is talking about? It is, and your sister should be advised accordingly. She should let her resentment out. It could be done in some direct or indirect way. An example of an indirect way would be her buying a punch bag and relieving her anger and resentment by violently assaulting the bag. An example of direct way would be to express her resentment addressing her boss. Notice, that the direct way doesn`t imply a performance of (HRHA). It isn`t implied even if her expression achieves the first-personal *practical* uptake by her boss of his wrongdoing; and it doesn`t do so for reasons discussed above.

We can safely conclude that holdings *accountable* can be suspended or given up for prolonged periods of time or even for good.

4.5.2.1 Jm-justification vs jf-justification

What about (HRHA) and a theoretical justification grounded in the free-will assumption? Is one needed? It depends. If we stick with Macnamara`s understanding of (HRHA) then it might turn out that no such theoretical justification is needed. The reader will have remembered that, according to Macnamara, the aim of (HRHA) is to induce the first-personal practical uptake of one`s wrongdoing. This is done by triggering an enforcement mechanism: a sanctioning behaviour or placing *burdens* onto the one guilty of the wrongdoing. Depending on various factors, the uptake will be achieved, partially achieved or not achieved at all.

Now, the enforcement mechanism does the enforcing by causing *pain*[154]. And, of course, causing pain to a sentient being calls for a justification. But what sort of justification? Here a distinction needs to be made between:

> *jm–justifications*, which are justifications related to the *moral* domain,
> and
> *jf–justifications,* which are justifications related to the *functional* domain.

Then we get the following conditionals:

i) *A* is *jm-justified* to cause pain to *B* only if *B deserves* it/it is *fair*[155] towards *B*.

ii) *A* is *jf-justified* to cause pain to *B* only if it, in some practical or functional sense, *benefits B*, and/or the community.

A murderer *deserves* the pain of imprisonment, and the judge is *jm-justified* to impose it on the murderer. A patient will benefit from a dental treatment, and the dentist is *jf-justified* to impose on the patient the pain associated with the treatment. I take it as immediately obvious that *jf-justifications* do not require grounding in the free will assumption. A dentist is *jf-justified* to cause the pain associated with the beneficial treatment even if the patient lacks any free will.

The issue of free will enters the stage only with *jm-justifications*: it doesn`t feel *right* to take the person who couldn`t have done otherwise as *deserving* the pain (of punishment).[156] This last point is, among others, a reminder of a claim made above (p. 77) that moral justifications have something to do essentially with *fairness*.

We have asked what sort of justification is required when imposing pain on others as part of (HRHA). On Macnamara`s understanding of what is involved in (HRHA), it looks like the sort of justification required is the *jf-justification*. You have expressed your

[154] The notion of pain is used here in a broad sense that covers both physical and psychological varieties of it.

[155] I will not distinguish between *fairness* and *desert* here, even though there are contexts where they behave semantically differently.

[156] That is not to say that *jm-justifications* require the free will assumption necessarily. There might be utilitarian readings of *desert* which will render *jm-justifications* free of the requirement. A murderer might *deserve* – under the utilitarian reading - his punishment because of the overall *benefits* it generates for him and society.

resentment or indignation in (HRDMA). You feel more is needed. The wrongdoer needs to be held *accountable,* that is: you aim at inducing the first personal uptake of the wrongdoing by causing pain. Notice that, on Macnamara`s understanding, you cause pain because it results in the uptake and *not* because the wrongdoer *deserves* it.[157] That is a crucial point. If *desert/fairness*[158] is not what motivates your (HRHA) then it can`t be that your (HRHA) stands in need of the *jm-justification*. Of course, this doesn`t, on its own, imply that it is the *jf-justification* that your (HRHA) stands in need of. For that to be true, we need to be able to read Macnamara`s notion of (HRHA) along functional lines, i.e. as referring to an act that is essentially motivated by generating practical *benefits*. This shouldn`t be too hard though. Why would one want to induce the first-personal practical uptake of a wrongdoing? Because, presumably, it is practically *beneficial* for the society if more – rather than less – people accomplish the uptake. Arguably, it is also *beneficial* for the wrongdoer herself as any such uptake will make her, morally speaking, a more mature person. Thus, it looks like once it is accepted that the essential aim of (HRHA) is to induce the uptake of wrongdoing, we find ourselves in the functional or practical territory where *jf-justifications* rule. And this kind of justification stands in no need of grounding in the free will assumption. This implies that (HRHA) – as understood by Macnamara - doesn`t require such grounding either.

4.5.3 (HRHA) and the free will assumption: contra Macnamara

Above, I promised that, once Macnamara`s account of (HRHA) is better understood, or just suitably corrected, it will turn out that our practice of holding another *accountable* actually *does* require grounding in the free will assumption. My strategy here is to show that in our practice of (HRHA) we *are,* as a matter of fact and contrary to Macnamara`s view, motivated by *desert/fairness.*[159] Moreover, this motivation by *desert/fairness* is, I shall claim, essential to our practice of (HRHA), unlike trying to induce the uptake, which is, I will argue, merely contingent to the practice and playing a different role in it.

I will ask the reader to consider two scenarios: a scenario in which the first personal uptakes are achievable via neurocognitive programming, a process that causes no discomfort, and a scenario in which the wrongdoer is, for some reason, incapable of performing the uptake. I shall call the scenarios *the neurocognitive scenario* and *the incapacitated*

[157] Macnamara doesn`t explicitly *deny* that one of the reasons you might have for inducing pain onto the wrongdoer could be that the wrongdoer *deserves* it. At the same time, nothing she says indicates that she would count *desert* as one of the reasons. Of course, there is a sense in which the target (the wrongdoer) of (HRHA) must have *deserved* your reactive attitude(s). The wrongdoer must have *deserved* it, minimally, in the sense of being *the cause* of the wrongdoing.

[158] See footnote 163.

[159] The phrase `motivated by desert` might sound somewhat ambiguous. I want it to be understood in the following way. When holding another accountable, one is motivated by desert if one seeks to give the person held accountable her dues.

wrongdoer respectively. The former scenario is meant to show that aiming to achieve the uptake is not a *sufficient* motivator behind our practice of holding others accountable. The latter is designed to answer the question whether the uptake is at least a *necessary* motivator of the practice. It will turn out that the uptake cannot be understood as a necessary motivator either. Yet, as it will become clear, the uptake plays an important role in the practice of (HRHA): it might be seen as a necessary condition of *fair* practice of (HRHA). This will clear the path towards claiming that the necessary motivator is *desert/fairness*. And, of course, if it is (preservation or restoration of) *desert/fairness* that motivates us in our treatment of a wrongdoer then that treatment calls for the *jm-justification*, i.e. the kind of justification that requires grounding in the free will assumption.

4.5.3.1 *The neurocognitive scenario*

Imagine that advances in neurocognitive programming make it possible to achieve the first-personal uptake of one`s wrongdoing via brain manipulation of the wrongdoer. The procedure causes no discomfort whatsoever to the wrongdoer. It has no unpleasant side- or after- effects. As a matter of fact, it is not uncommon that the procedure is felt as mildly pleasant and relaxing. Also, undergoing the procedure will result in an above-average probability that the wrongdoer won`t resort to any similar kind of wrongdoing ever again.

Now, welcome Brian to the story. Brian has committed a murder and is to be held responsible for the gruesome deed. Brian has already been subjected to a vast array of reactive attitudes – such as anger, resentment, indignation, etc. – via which various people held him responsible in the sense of *deep moral appraisal* (HRDMA) and, to some extent, also in the sense of *accountability* (HRHA).[160] Now he finds himself in front of a judge and about to be held *fully*[161] accountable (HRHA) for his crime.

The reader will have remembered that according to Macnamara the aim of holding others accountable for their wrongdoing is to induce the uptake. We are *successful* in our

[160] Highly likely, Brian has already been subjected to reactive attitudes that involved holdings accountable. At least some of his friends and relatives, presumably, must have approached him with the aim of inducing the first personal uptake. Yet, I believe that Macnamara would agree that the uptake of something as serious as a murder requires a correspondingly serious enforcement mechanism to be employed. Thus, Brian is taken to a courtroom.

[161] Why do I talk about *full* accountability here? The logic of the scenario seems to put Macnamara`s notion of holding accountable under more conceptual stress that it might be able to withstand. The logic of the scenario gives rise to questions that Macnamara herself doesn`t ask and, of course, doesn`t answer. For instance: Should the *force* in the enforcement mechanism employed to achieve the uptake somehow reflect or correspond to the seriousness of the wrongdoing? If yes, why? This question will be, in a sense, answered below. For now, I will assume as intuitively obvious that the *force* of the enforcement mechanism that is at the core of holding others accountable should in some sense reflect or correspond to the moral seriousness of the wrongdoing.

holding another accountable if the uptake is achieved by the wrongdoer. The judge then approaches the case of the murder under the following assumptions:

A1: My job here today is to hold Brian (fully) accountable for his crime.

A2: To hold another (fully) accountable for their crime is to (successfully) induce the first- personal uptake of the wrongdoing.

A3: The widely available procedure of neurocognitive programming offers 100% success rate in achieving the uptake.

Constrained by the logic of the three assumptions, the judge sentences Brian to a session of neurocognitive programming designed to induce the uptake, and moves onto another case.

Now, is it a happy ending? Does it feel *right* that a human being gets murdered and *all* that is deemed appropriate is, (a), to manipulate the murderer`s consciousness into (fully) realizing what a horrible crime he has done[162] and, (b), making it – via the manipulation – improbable that the murderer will murder again? I will hope that majority of readers will join me in feeling that this wouldn`t be *right*.

4.5.3.1.1 An objection

I shall get ahead of my argument a bit and suggest that the verdict doesn`t feel *right* because Brian hasn`t got what he *deserves*. And what he deserves is *pain* in some form.[163] I admit that I do not know how to argue for this claim apart from asking what else could explain the feeling. The question, however, invites the following objection: True, it doesn`t feel *right* to limit our response to Brian`s horrible crime to making him undergo the neurocognitive procedure. *Pain* in some form – and of a degree corresponding to the seriousness of his crime – needs to be imposed on Brian. However, it is not because Brian *deserves* the pain but because his suffering the pain works as a *deterrent* to other would-be murderers.

The objection removes *desert* from the picture of holding accountable, (HRHA), and pushes it back into the functional domain. There, in the functional domain, it won`t call for a justification grounded in the free will assumption. To counter the objection, *the neurocognitive scenario* needs to be modified: Imagine that the would-be offenders in Brian`s community (or society as a whole) don`t, for some reason, get deterred by the pain

[162] Below, I will argue that there is a degree of pain that is intrinsic to achieving the uptake as such. Thus, those readers who will agree with me that, indeed, the uptake is always, to some degree, painful, might feel tempted to reply that it feels quite all right to limit the whole punishment to just manipulating Brian`s brain into realizing the uptake. Such a reply is fine with me as it is perfectly compatible with my target claim here that what we are ultimately looking for when evaluating the justness of a sanctioning behaviour is whether (the right degree of) *pain* has been imposed.

[163] This is not to say that inducing pain in some form is *all* that is to be done. There is no reason not to combine it with the neurocognitive session, for instance.

of punishment imposed on those who commit a crime and get caught. In Brian`s community, statistics and experiments have shown conclusively that no matter how severe the punishment is the crime rate stays unaffected. Various explanations for the unexpected fact are proposed: there is a neurological anomaly regarding the criminal mind that a criminal always and sincerely believes that she won`t be caught; people`s imaginations have been oversaturated by the violent images the media exposes them to, and, as a result, they are unable to relate to the pain of people they don`t know; or perhaps, it is not the violent images that make them incapable of relating to the pain that should deter them, but a common ingredient in processed food they eat every day that causes the incapacitation. Whatever the reason turns out to be, it is the case that in Brian`s community imposing pain is a zero deterrent.[164]

Against the background of the modified scenario, I shall ask the same question I asked against the background of the original one: does it feel *right* that a human being gets murdered and *all* that is deemed appropriate is, (a), to manipulate the murderer`s consciousness into (fully) realizing what a horrible crime its bearer has committed and, (b), making it – via the manipulation – improbable that the murderer will murder again? Now, it can`t be answered: true, it doesn`t feel *right* but that feeling has nothing to do with *desert* and everything to do with the failure to establish or reaffirm a *deterrent*. Such an answer cannot be given because the modified scenario explicitly rules out the possibility that imposing pain as part of punishment can serve as a deterrent in Brian`s community. Thus, we are back to my claim that the verdict doesn`t feel *right* because that`s not what Brian *deserves* (or what is, in some sense, *fair*); and what he deserves is (corresponding) *pain* in some form. And once more I shall admit that although I do not know how to argue for the claim, I am, at the same time, unable to see what else could explain the feeling.

I will conclude that *the neurocognitive scenario* shows that when holding others accountable, (HRHA), the aim of achieving the first-personal uptake of a wrongdoing cannot be what *sufficiently* (if at all) motivates the practice. The uptake, i.e. the goal of the practice, has been achieved and yet we find the verdict unsatisfactory. Something else or more is needed. At the same time, the scenario points towards *desert/fairness*[165] as playing an

[164] This modification to the original *neurocognitive scenario* shouldn`t be seen as too fantastic. A deterrence effect of cruel punishments is rather questionable. For instance, the death penalty – which is a punishment that one would expect to be a strong deterrent – doesn`t seem to be such at all. Available data show conclusively that the death penalty doesn`t serve as a deterrent. See for instance John Lamperti, *Does Capital Punishment Deter Murder? A Brief Look at the Evidence*, 2010 < https://math.dartmouth.edu/~lamperti/my%20DP%20paper,%20current%20edit.htm> [accessed 17 September 2019]. Lamperti concludes his survey and evaluation of the available evidence for deterrence with the following words: `[T]he data which now exist show no correlation between the existence of capital punishment and lower rates of capital crime`.

[165] The reader is asked to restrict her reading of `desert` to contexts where it relates to *blameworthy* conduct only. It is not clear that giving another what they deserve requires justification grounded in

important, perhaps even an essential, role as a motivator in the practice of (HRHA). And a motivation by desert calls for a justification that is grounded in the free will assumption. This transfers onto the whole practice of (HRHA) making it stand in need of a justification grounded in the free will assumption.

We have established that achieving the uptake cannot be the sufficient motivator behind our practice of (HRHA). At the same time, it seems to be clear that achieving the uptake plays an important role in the practice. The question is what exactly the role is. Is it at least a *necessary* motivator? Is it perhaps *necessary* for the practice not as a motivator but in some other sense? The following scenario and its discussion are meant to answer these questions.

4.5.3.2 *The incapacitated wrongdoer*

Brian`s father happens to be a neuroscientist capable of manipulating psychological and emotional responses of people via neuro-programming. One of the things he is capable of is programming someone`s brain in such a way that the person permanently and irreversibly loses the capacity to achieve the first-personal practical uptake of a wrongdoing. Brian`s father is an educated man. He knows that the legal system in his society is explicitly grounded in Macnamara`s ethics. That is, he knows that the legal system takes punishment to be essentially and solely about achieving the uptake. Long-term prison sentences are agreed to be an efficient way of achieving the uptake. Despite Brian having committed a murder, his father still loves him and cannot come to terms with the idea of his son spending a decade or two behind bars. He devises a desperate plan. Via neuro-manipulation of his son`s brain he permanently and irreversibly deprives his son of the capacity to achieve the uptake. He also makes sure to `shut down` all the neural paths in Brian`s brain associated with violent behaviour making it extremely improbable that Brian would intentionally commit a violent crime ever again. The neuro-programming procedure has been recorded in full length, brain scans taken, to be used as evidence that Brian has been incapacitated regarding achievability of the uptake. In court, Brian`s father argues:

`Your Honour, let`s take *achieving A* to be the only reason for *doing B.* It turns out that *A* is, as a matter of fact, *unachievable.* The unachievability of *A* removes, I claim, the only reason for *doing B.* Applied to my son`s case: Our legal system takes achieving the first-personal practical uptake of a wrongdoing as the essential and sole, that is, the only, reason for imposing a punishment. This sole reason – inducing the uptake – is unachievable in the case

the free will assumption in the context of *praiseworthy* conduct. My view is that giving one her dues in response to her praiseworthy conduct doesn`t stand in need of a justification grounded in the free will assumption.

of my son as evidenced by the neuro-programming procedure recording, brain scans and testimonies of my neuroscientific colleagues. Thus, there is no reason to punish my son.`

There is a pause. The judge thinks about the argument. It occurs to him that it is unclear in what sense it is correct to claim that a reason for doing something is `removed` or stops being there just because that something cannot be done. He soon stops worrying about this subtlety though. He realizes that what the argument of Brian`s father shows, at the very least, is that it would be totally *pointless* to punish Brian. And the judge would hate to be associated with *pointless* decisions. He lets Brian go free.

Now, we don`t know enough about the society that Brian and the judge live in. The moral intuitions of its members might be such that they would not find anything `out of tune` with the judge`s decision. If anything, perhaps, they would applaud the judge`s commitment to ground his decisions in sound reasoning. However, I am confident that at least some of my readers will find the judge`s decision somewhat troubling. Brian wasn`t incapacitated in any relevant sense at the time of the murder, or, at least, there is nothing in the scenario indicating so. He, let`s assume, intended, planned, committed and tried to cover a crime of murder. And it can`t feel *right* to let him go unpunished just because, at some point *after* the crime, he lost the capacity to achieve the uptake. Agreed? Unfortunately, there are issues that complicate the picture.

It could be objected that how one feels about the verdict depends on what precisely (and how much) gets lost with the loss of the capacity to achieve the uptake. Let`s consider a case that might seem rather extreme at first. Imagine that together with losing the ability to achieve the uptake, Brian also loses, (a), the ability to feel any empathy whatsoever towards another sentient being and, (b), a related ability to understand that there is anything wrong with hurting someone. This shouldn`t be perceived as too fantastic and arbitrary a scenario. On the contrary, the psychology of the uptake seems to overlap substantially with the psychology of (a) and (b). The reader will recall that, according to Macnamara, to induce the uptake is `to get the wrongdoer to acknowledge her wrongdoing, feel remorse, apologize, make amends, and commit to doing right in the future`[166]. Thus, the person incapable of the uptake is incapable of acknowledging her wrongdoing, incapable of feeling remorse, incapable of apologizing and incapable of committing to doing right in the future. Of course, the person might still be able to *pretend* to acknowledge, feel, apologize and commit, but that is not what the uptake is meant to be. Now, aren`t the inability to acknowledge one`s wrongdoing, the inability to apologize and the inability to commit to doing right in the future all cases of (b)? And isn`t the inability to feel remorse a case of (a)? They are, and we need

[166] Macnamara, `Holding others responsible`, p.90.

to adjust our understanding of Brian`s incapacitation accordingly: Brian is not only unable to achieve the uptake, he is, moreover and consequently, unable to comprehend any legitimacy of the punishment. He can`t but see it as an irrational and bizarre custom and himself as terribly unlucky to be subjected to it. How does it feel now when it comes to punishing Brian?

I suspect that at this point the opinions will get divided. The issue here is not dissimilar to that concerning the moral agency of psychopaths. Due to their inability to feel empathy towards others, psychopaths seem to be unable to recognize that hurting others is wrong. This inability to recognize right from wrong makes them non-members of moral community. Consequently, the binding oughts of the community can`t reach them; i.e. they cannot be a target of holding accountable. Some philosophers find the conclusion troubling. T. M. Scanlon, for instance, argues that the lack of empathy doesn`t translate into an inability to recognize rights from wrongs.[167] Gary Watson disagrees: `[A]nyone who is incapable of recognizing the interests of others as making valid claims on her is incapable of grasping and responding to moral requirements` and `no one who is morally incompetent in this way is fit to be held morally responsible`.[168] Another dissenter here would, presumably, be John Rawls, who argues that `the duty of justice is owed to those [only] who are capable of a sense of justice`[169], which seems to be something that psychopaths are incapable of. Some could insist that this `argument from psychopathy` doesn`t really bite here because it is different from Brian`s case in an important respect. A psychopath has always been a psychopath while Brian has become one only after the crime. And it seems, at least on the face of it, that his ability to recognize right from wrong at the time of the crime makes the case very different from that of someone who wasn`t able of such recognition at that time. These are intriguing issues. I shall, however, not venture into discussing them. They have been briefly mentioned to (a) show that the issue is far from clear and (b) to set a background against which the role of the uptake in the practice of (HRHA) can be clarified.

4.5.3.3 The motivator: uptake vs fairness

At various points, when discussing the scenarios, we asked whether a verdict felt *right*. The question is ambiguous in an important sense. Did it (or not) feel right because the scenario `removed` the motivator, or because it exposed a deficit in *desert/fairness*? Giving the former

[167] Thomas M. Scanlon, *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press, 1998), ch.5.

[168] Gary Watson, `The Trouble with Psychopaths`, in R. Jay Wallace, Rahul Kumar, and Samuel Freeman (eds), *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon* (Oxford: OUP, 2011), pp.308-24 (p.309).

[169] John Rawls, `The Sense of Justice`, *The Philosophical Review*, 72, no. 3 (1963), pp.281-305 (p.281).

as the answer would imply that the uptake is a necessary motivator while giving the latter as the answer would imply that although the uptake is not a necessary motivator, it still plays an important role of a necessary condition for *desert/fairness* in the practice of (HRHA).

The question is a bit tricky as the two possible reasons for the verdict not feeling right could be seen as closely related. The relation is that punishing someone when the punishment cannot be motivated seems *unfair*. Thus, removing the motivator amounts to causing a deficit in *fairness*. This, of course, turns on the assumption that *desert/fairness* is what ultimately motivates our practice of (HRHA). This assumption is what I am arguing towards. It is, however, not what Macnamara seems to endorse. She takes the uptake to be not only a necessary motivator but the *sufficient* or *ultimate* one.

Let`s have a look at *the neurocognitive scenario* again. In the scenario, we test Macnamara`s claim that the sufficient motivator behind our practice of (HRHA) is to achieve the uptake of a wrongdoing. I shall assume that the scenario has shown that we would wish to punish Brian even after the motivator (the uptake) has been removed. This implies that it must be something *else* that motivates our practice of (HRHA). I submit that what *ultimately* and *sufficiently* motivates us is (the normative force of) *desert/fairness*. However, as *the incapacitated wrongdoer* scenario shows, it won`t be that easy to dismiss Macnamara`s concept of the first-personal uptake and the role it plays in the practice of (HRHA). First, it seems obvious that often we really wish `to get the wrongdoer to acknowledge her wrongdoing, feel remorse, apologize, make amends, and commit to doing right in the future`[170], i.e. it seems obvious that often what motivates us is to induce the uptake. Second, our moral intuition seems to lose its bearing when it comes to punishing a wrongdoer that is uncapable of the uptake. In other words, once `the uptake` is removed from the picture, it gets rather unclear whether a punishment is appropriate or not. This strongly suggests that the uptake plays an important role in the practice. What role though?

To answer this question, I shall propose and argue for the following claim: The uptake is not (even) a necessary *motivator*. A necessary (and, in fact, the sufficient) motivator is desert/fairness. The uptake is, at best, a necessary *condition* for desert/fairness to obtain or be preserved. In what follows, I shall employ the claim to answer all those questions that rose in connection with the scenarios. I hope to make the claim convincing via showing its explanatory force in answering the questions.

Let me first deal with perhaps the most conspicuous inconsistency related to the claim. Towards the end of the previous paragraph, I concede that `often what motivates us is to induce the uptake`. Yet, a few sentences below, I claim that the uptake is not a (necessary) motivator at all. An easy reply to the perceived inconsistency would be that something being

[170] Macnamara, `Holding others responsible`, p.90.

the case *often* doesn`t translate into something being the case *necessarily*. This answer won`t do. It would be easy to think of scenarios where inducing the uptake is *all* we wish to do in response to someone`s wrongdoing. A wife reproaching her husband who has, once again, come home drunk. Surely, *all*[171] she is after is inducing the uptake. Thus, while the uptake is perhaps not a necessary motivator for the practice of (HRHA) as a *kind* of holding responsible, it is, at the same time, hard to deny that often it is, in fact, *all* that motivates particular instances of the practice. This fact is, of course, something that my account of (HRHA) has to be able to explain. If it is the case that often *all* that motivates our practice of holding others accountable is inducing the uptake then, clearly, my claim that the practice is essentially motivated by *desert/fairness* can`t be true. So, what is happening here?

According to Macnamara, as discussed above, the uptake is induced via imposing (broadly speaking) *pain* onto the wrongdoer. Thus, the pain is a means of achieving the uptake; it is something that is *external* to the uptake. This picture of how pain and the uptake relate to each other is plausible and captures correctly, I believe, what happens, among others, when we attempt to hold others accountable. There is, however, something here that might pass unnoticed. The uptake *itself* involves a degree of pain; that is, there is a degree of pain that is *internal* to the uptake. Let`s have a look again at what is involved in the uptake. The reader will have remembered that, according to Macnamara, to induce the uptake is `to get the wrongdoer to acknowledge her wrongdoing, feel remorse, apologize, make amends, and commit to doing right in the future`.[172] To acknowledge one`s wrongdoing is a humiliating and/or humbling experience. It is a kind of experience that we, typically, wish to avoid. We wish to avoid it because it is painful. The same applies to the act of apologizing. Feeling remorse is an emotion that too is painful, although for reasons different from those associated with acknowledging wrongdoing and apologizing. Feeling remorse is not a humiliating experience. Humiliation about something presupposes awareness of other`s awareness of that something. Feeling remorse is essentially a private emotion and as such doesn`t depend on other`s being aware of anything. Feeling remorse is perhaps a humbling experience and to this extent it is an experience that is painful. Even if not seen as humbling, it is clearly an experience that we normally wish to avoid. This, again, strongly indicates that it is an experience that is, broadly speaking, painful.

Making amends and committing to doing right in the future are requirements involved in the uptake that won`t reduce to pain as easy, if at all, as the previous ones. Making amends might be experienced as painful in some cases and agreeable in others depending, perhaps,

[171] `All` in terms of motivation. Of course, the wife`s reproaching might – and very probably will – *involve* expressing emotions constitutive of *deep moral appraisal*. But that`s not what motivates her if what she does is trying to hold her husband *accountable*.
[172] Macnamara, `Holding others responsible`, p.90.

on the maturity of one`s conscience; i.e. a person tormented by guilt (that is, someone capable of having guilty conscience) will often find relief in making amends. This fact suggests that making amends is not essentially about pain. There are two reasons why this doesn`t present a problem for my claim that the uptake itself involves a degree of pain. First, the truth of the claim doesn`t require *all* aspects that constitute the uptake to be reducible to pain. An apple is rotten even if a part of it is not. Second, the reader will remember that the reason I am trying to convince her that the uptake itself involves a degree of pain is that it will allow me to claim that it is desert/fairness that motivates our practice of (HRHA). Now, making amends seems to be an act that can be about desert/fairness – in the sense that it attempts to restore that which got disrupted in the act of wrongdoing[173] – without, at least sometimes, being so *in virtue of* imposing pain. A person who has damaged a public property will be ordered to pay for its repair which the person might find nearly totally painless if she is sufficiently rich. This particular instance of making amends might not be convincing because perhaps making amends in the financial domain is something rather different from making amends in the moral domain. Making amends in the financial domain might be translatable into making amends in the moral domain if pain is imposed but that`s not what always happens.[174] Macnamara`s definition of the uptake is ambiguous here and as such is consistent with both understandings of `making amends`. Be that as it may, there is no need to establish conclusively what exactly is the role of making amends in the uptake as it won`t affect the truth of my claim that the uptake involves a degree of pain.

The last constituent of the uptake mentioned by Macnamara is committing to doing right in the future. This is an odd one here. In morally neutral contexts, the committing to doing right in the future can be seen as involving, at best, only a very minimal degree of pain. It will be seen as such by those who feel that any commitment to future action is an act of voluntary self-constraint that involves a degree of psychological discomfort or pain. Others might feel that there is nothing discomforting or painful in the act of committing to doing right in the future in morally neutral contexts. We don`t need to take sides here because what interests us is not the act of committing in the morally neutral contexts but in those that are morally significant (and recall: the context of the uptake is the context of a wrongdoing). Here, it seems to me, the act of committing to doing right in the future involves a stronger degree of pain (compared to the degree it involves in the morally neutral contexts) in virtue

[173] Below, I shall explain *fairness* in terms of the notion of *moral equilibrium*. A wrongdoing disrupts the equilibrium. Restoring the equilibrium is what motivates our practice of (HRHA), and the equilibrium gets restored in an (successful) instance of (HRHA).

[174] In Switzerland, the fine for speeding varies depending on the financial status of the offender. The richer the offender is the higher the fine. This fact could be interpreted along the utilitarian line or along the desert/fairness line (or combined). The utilitarian interpretation would be that in this way the force of deterrence extends to all strata of society. The desert/fairness interpretation would be that in this way pain can be induced (and fairness restored) in all, or nearly all, cases.

of it being, in fact, an act that is meant to be humbling. When one is induced to perform an explicit commitment it seems to be, psychologically speaking, an experience that is very similar to that of acknowledging wrongdoing or apologizing. It is humbling, perhaps even humiliating in some cases, and, as such, a painful experience.

The above is, admittedly, a very sketchy analysis. It will suffice though as nothing of importance hinges on the status of the acts of making amends and committing to doing right in the future regarding whether or not the two acts involve a degree of pain. My claim that the uptake involves a degree of pain is fully sustained in virtue of the acts of acknowledgement of wrongdoing and apologizing being clearly acts that are painful.

I will take it as established that the uptake involves a certain degree of pain that is internal to it. How does it help us with the inconsistency, i.e. with claiming, on one hand, that (HRHA) is *essentially* motivated by desert/fairness and admitting, on the other, that the uptake is often *all* that motivates instances of (HRHA)? Having established that the uptake itself is intrinsically painful allows me to claim that the motivation by the uptake boils down to the motivation by desert/fairness. The reader will have remembered that the modified neurocognitive scenario (the one that explicitly rules out utilitarian motivation) strongly suggests that what motivates our practice of (HRHA) is restoring fairness (i.e. moral equilibrium) disrupted by a wrongdoing, via imposing pain the degree of which corresponds to the seriousness of the wrongdoing. Now, if, as has been established, the uptake is intrinsically painful, then the fact that it is often all that motivates (HRHA) cannot be understood as contradicting my claim that (HRHA) is essentially motivated by desert/fairness. The intrinsic painfulness of the uptake allows us to take it as an instance of restoring the moral equilibrium and allows me to keep claiming that our practice of (HRHA) is essentially motivated by desert/fairness.[175]

We have established that being motivated by the uptake is – in virtue of the uptake being intrinsically painful – being motivated by desert/fairness. In other words, the uptake is

[175] The essential painfulness of the uptake allows for the following `Macnamarian` reading of the neurocognitive scenario: It could be plausibly argued that any robust uptake will involve feelings of guilt accompanied by a degree of pain that will correspond to the seriousness of the wrongdoing. If we tie the uptake with required levels of pain in this way, we can protect Macnamara`s account from the force of the neurocognitive scenario. A Macnamarian could keep insisting that it is the uptake that motivates our practice of holding others accountable and we wouldn`t be able to object on the grounds that a *deserved* degree of pain hasn`t been induced. Such a Macnamarian account would tick all the boxes. I would find such an account very attractive as it would capture the following two things that I believe are essential to what motivates the practice of holding accountable. First, it would correctly capture the feeling that the wrongdoer must be burdened with a corresponding degree of pain. Second, it would correctly capture our desire that the wrongdoer fully realizes the wickedness of her actions and broadens her moral consciousness. Be it as it may, my argument in this chapter doesn`t require that desert/fairness (that calls for imposition of pain onto a wrongdoer) is a *sufficient* motivator. All that is needed is that it is a *necessary* one; and that`s what has been showed by the neurocognitive scenario.

a motivator only to the extent to which it is painful.[176] The discussion above of *the incapacitated wrongdoer* scenario raised some doubts about the moral appropriateness of punishing a wrongdoer who is incapable of achieving the uptake. It could be felt as unfair to punish a wrongdoer who – as a consequence of his inability to achieve the uptake - cannot comprehend the legitimacy of the punishment, who can`t but see it as an irrational and bizarre custom and himself as terribly unlucky to be subjected to it. Those who feel so might then wish to list the ability to achieve the uptake as a necessary *condition* of a *fair* punishment (or of a *fair* exercise of HRHA). And that is perfectly compatible with my claim that our practice of (HRHA) is essentially motivated by desert/fairness. The logic in the background is this. Restoring the moral equilibrium (the fairness) that has been disrupted by a wrongdoing can be done only by a punishment (or an exercise of HRHA) that is itself *fair*. An unfair punishment disrupts the moral equilibrium even more. (More about this later)

Let me briefly sum up the role of the uptake in our practice of (HRHA). *The neurocognitive scenario* shows that the uptake cannot be what necessarily motivates the practice. The modified neurocognitive scenario points to *desert/fairness* as the only plausible motivator of the practice. *The incapacitated wrongdoer* scenario is meant to shed some light on the role of the uptake in the practice of (HRHA). Is the uptake all that motivates the practice at least in some instances? Or is it perhaps crucial for the practice not as a motivator but in some other sense? It turns out that (a) those instances of (HRHA) that look like they are motivated solely by the uptake are in fact motivated by desert/fairness and (b) the uptake can be understood as, at best, a necessary *condition* of successful restoration of moral equilibrium, i.e. of fairness.

4.5.4 Conclusion: holding accountable calls for the free will assumption

Above (p.88), I claim that `[t]he issue of free will enters the stage only with *jm-justifications*: it doesn`t feel *right* to take the person who couldn`t have done otherwise as *deserving* the pain (of punishment).` The claim is, in other words, that it wouldn`t be *fair* to impose pain on a wrongdoer for a wrongdoing that he couldn`t have avoided doing, i.e. on a wrongdoer who lacks free will. When holding others accountable, we impose pain. Holding others accountable is motivated by desert/fairness. Claiming that imposing pain on a wrongdoer in the act of holding accountable wouldn`t be fair if he lacks free will implies that the practice

[176] The issue here is a bit slippery and the reader might still feel that often, as a matter of fact, all that motivates her holdings of others accountable is achieving the uptake. I agree. However, I insist that all those cases where achieving the uptake is the sole motivator are cases in which the seriousness of the relevant wrongdoing is relatively low and can be counterbalanced by the relatively limited degree of pain that is intrinsic to the uptake. In those cases in which the seriousness of the relevant wrongdoing (think of murder, for instance) cannot be counterbalanced by the pain that is intrinsic to the uptake, the uptake will be complemented by imposing additional burden.

of (HRHA) - the practice of holding accountable – calls for the free will assumption. And that is what I have been arguing towards.

4.6 Fairness: three claims

Now, when the circle of the argument behind the claim is closed, I can go back and look again at some of the claims that the argument relies on. The plausibility of an argument depends on the plausibility of its individual claims and I feel that more should be said about some of the claims to boost the overall plausibility of the conclusion.

The claims in need of further discussion are the following:

B1: Imposition of pain is justified if *deserved/fair*.

B2: Imposition of pain on a wrongdoer is deserved/fair only if the wrongdoer has free will.

B3: Holding others accountable is motivated by desert/fairness.

Let me start the discussion of (B1) and (B2) with a little reminder of why they need a discussion at all. I have been defending a claim – contra Strawson – that there is a kind of holding morally responsible that calls for a justification grounded in the free will assumption. The kind of holding morally responsible that calls for such a justification is the kind that Macnamara calls *holding accountable*. It is an empirical fact that we often engage in this kind of practice, i.e. the existence of this practice is uncontroversial. The controversial bit is whether holding accountable calls for a justification grounded in the free will assumption. This is where the plausibility o (B1) and (B2) becomes important. Holding accountable is a kind of practice that essentially involves the imposition of pain onto another. Imposing pain onto another is – according to (B1) – justified only if deserved/fair, which is something that is – according to (B2) - deserved/fair only if the wrongdoer possesses free will. Thus, the plausibility of the claim that holding accountable calls for a justification grounded in the free will assumption depends on the plausibility of (B1) and (B2). (B3) is a claim that has been endorsed above when discussing the *neurocognitive scenario*. As formulated, the claim needs some unpacking. The unpacking of (B3) will follow the discussion of (B1) and (B2).

4.6.1 Imposition of pain is justified if deserved/fair

(B1): Imposition of pain is justified if *deserved/fair*. Above (p. 88), I distinguish between, broadly speaking, utilitarian justifications (*jf-justifications*) and moral justifications (*jm-justifications*). (B1) is a claim about imposing pain as a response to a morally wrong conduct. It is such a claim because it is made in the context of (HRHA), which is a context that involves morally significant behaviour and our responses to it. Thus, the justification mentioned in the claim is to be understood as of a moral kind, i.e. as the *jm-justification*. I have, until now, relied on an immediate intuitive plausibility (and clarity) of the claim and explicitly or implicitly appealed to it without any defence of it. The time has come to say more about it.

One of the problems that interests moral philosophers is the problem of what justifies punishment; i.e. the problem of what makes it morally permissible to punish another.[177] This problem and its discussion is strongly analogous to what can be expected to be involved in the discussion of (B1). The most general and uncontroversial definition of punishment is: imposition of burden of some sort in response to a wrongdoing. I have been treating imposition of pain and imposition of burden interchangeably throughout the chapter. There might be some differences between the two. However, whatever the differences might be, it is clear that pain *is* burdensome. Thus, imposing pain is a case of imposing burden. Consequently, whatever one might say about justifiability of punishment will apply to justifiability of imposing pain, which is my concern regarding (B1).

The philosophical discussions of the permissibility of punishment reduce, ultimately, to two positions: consequentialism and retributivism. According to consequentialism, the rightness or wrongness of something – an action, a rule for action, or an institution – is determined solely by the consequences. An action, such as imposing pain, or generally any punishment, will be justified if it produces the right kind of overall consequences. Consequentialism thus takes punishment as a *means* towards some valuable end. Retributivism, on the other hand, takes punishment as, in a sense, a goal in itself, as something that is intrinsically a right response to a crime. And, according to retributivism, punishment is intrinsically right and justified because it is *deserved*.

Now, clearly, (B1) – the claim I am defending here - is a retributivist claim. Retributivism competes with consequentialism regarding the problem of justification of punishment. This raises the following question: does the existence and logic of the consequentialist position pose a threat to the truth of (B1)? The answer to this question depends on whether the two positions are mutually exclusive or whether there is a way to combine them into a single account. If there is a way to combine them into a single account, then (B1) could be accepted as stating a necessary – not sufficient – condition in such a combined account (consequentialism would then provide the other necessary condition(s) to make the account complete). If the positions are mutually exclusive then, of course, the truth of consequentialism invalidates (B1). Let me take the hard path and presume that the

[177] There is a vast literature on the issue of moral justification of punishment. The proposed accounts dealing with the issue can be divided into, roughly, three groups: consequentialist, deontological and mixed. Perhaps the best collection of essays covering all the three groups of accounts is Michael Tonry (ed.), *Why Punish? How Much? A Reader on Punishment* (Oxford: Oxford University Press, 2011).

positions are mutually exclusive. That would imply that if consequentialism is true then (B1) is false. So, is consequentialism true? And if it is, do we have to abandon (B1)?

I shall not dwell much on the first question because even if the answer turned out to be affirmative there is a strong reason, I will claim, to stick with (B1). I confess that I find the consequentialist accounts of permissibility of punishment extremely implausible and am unable to force myself to take them seriously. The readers who are able to take them seriously are referred to the neurocognitive scenario, which can be understood as a thought experiment designed to counter consequentialist accounts in general, to test their intuitions. As I said above, I won`t argue against consequentialism in any more detailed way for a simple reason: I don`t have to. There is – in the context of our discussion of how holding accountable relates to the free will assumption – a serious reason to steer away from consequentialism and accept (B1).

The reason is that a consequentialist won`t see free will – i.e. its presence or absence in the agency of a wrongdoer – as playing any role in the moral appropriateness of punishing. For a consequentialist, the rightness of punishing is fully determined by the value of its consequences. Whether the wrongdoer lacks or possesses free will at the time of the wrongdoing that he is being punished for is something that cannot be construed as a consequence of the punishment simply because it precedes the punishment, i.e. it precedes that in relation to which the envisaged benefits can be seen as consequences. In other words, a consequence is – by definition – that which comes *after* something, not before.

Now, why is it a serious reason to steer away from consequentialism in the context of discussing how holding accountable relates to the free will assumption? Let me remind the reader of an important point here. My account of freedom in agency won`t be of any use in grounding justifications of holding others morally responsible. At the start of this chapter, I concede that that is not an attractive feature of the account. The implicit assumption throughout the whole chapter has been that the presence or absence of free will indeed plays a crucial role in the justification of our practice of holding others responsible. This is, however, an assumption that a consequentialist must refuse. A consequentialist is unable to enter the framework of the discussion as he doesn`t accept one of the framework`s crucial assumptions. Free will is not a consequence, and a consequentialist will (have to) ignore it.

A consequentialist will not, therefore, find my account of freedom unattractive (and if he does then it will be for some other reasons) and will skip this chapter. The crucial point here is this: *whether aware of it or not, any philosopher working under the assumption that free will is crucial for grounding our practice of holding others responsible must - on pain of being inconsistent - be a retributivist*. Thus, once we have accepted the assumption – as we had at the beginning of the chapter – we are bound to accept (B1).

4.6.2 Imposition of pain on a wrongdoer is deserved/fair only if the wrongdoer has free will

(B2): Imposition of pain on a wrongdoer is *deserved/fair* only if the wrongdoer has free will. Why is imposition of pain *deserved/fair* only if the wrongdoer has free will? I don`t know the answer to this question and I suspect there might not be one. The vast majority of literature on free will and moral responsibility turns on the assumption – the assumption that is analogous to (B2) - that it would be inappropriate to hold others morally responsible had they no free will. This assumption is frequently appealed to but never argued for.

There are accounts of moral responsibility that, on the face of it, seem to do away with free will (Harry Frankfurt`s and his followers`) but when one looks closer it turns out that those accounts work with a rather narrow definition of free will. Working with a narrow definition of free will allows those accounts to generate attractive claims and, at the same time, remain intuitively plausible in virtue of leaving those aspects of free will within their accounts that had been left out of the narrow definition.[178] In other words, even the accounts that seem to be able to ground our practice of holding others responsible without an appeal to free will, turn out – when scrutinized – to rely on some version of it. So why is it that philosophers don`t seem to be bothered about answering the question why it would be inappropriate to hold morally responsible those who lack free will? The only answer I am able to give is that the philosophers working on moral responsibility take the claim that only agents possessing free will can be justifiably held morally responsible to be an intuitively primitive claim, i.e. a claim that cannot be construed as the conclusion of a logical arrangement of some other claims. In other words, they take it to be a claim that captures a *foundational* moral intuition.

4.6.2.1 The sense of justice

I could perhaps rest my case at this point and move onto discussion of (B3). I wish to say a bit more about it though. The following remarks are, I believe, contextually useful (for later purposes) generalizations that can be associated with (B2). Also, the remarks will make the picture of how (HRHA) works richer and more plausible.

[178] Thus, Harry Frankfurt first defines free will as availability of alternatives. Then he conceives of a scenario in which the agent doesn`t have – in a certain sense - any available alternatives. The scenario is conceived in such a way that it doesn`t seem to be inappropriate to hold the agent responsible even though she has – in a certain sense – no available alternatives. Harry Frankfurt concludes that free will is not needed to hold responsible. Frankfurt`s scenario, however, leaves it open whether the past of the agent has been determined or not because the alternatives in the scenario have been removed not because of the determination running from the past but because of the determination running from the future. This trick, if I may say so, fools our moral intuition because there seems to be a robust sense in which the agent is free even though she hasn`t got – in a certain sense – any available alternatives. For the scenario, see Harry Frankfurt, "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy*, 66, no. 23 (1969), pp.829–839.

One of the reasons for (B2) being a primitive claim is that it employs a primitive concept of *fairness*. It is a primitive concept in the sense that (a) it cannot be defined but only acquired through practice, and (b) it has an axiomatic status in any moral discourse.[179] And while *fairness* cannot be defined, it can be made more *transparent* if we can tie it to a synonym that is (more) transparent. I propose *moral equilibrium* as such a (more) transparent synonym to *fairness*.[180] The notion of moral equilibrium exposes some important aspects of fairness that might otherwise pass unnoticed: fairness is a moral property; fairness has something to do with careful balancing of pros and cons; fairness doesn`t come in degrees: like equilibrium, it either obtains or not, nothing in between.[181]

Tying *fairness* to *moral equilibrium* allows the following straightforward rendering of (B2): Imposition of pain on a wrongdoer disturbs moral equilibrium if the wrongdoer is not a free agent. How do we know that the moral equilibrium is (or would be) out of balance in the case of (B2), and in other cases generally? (B2) is a primitive claim, thus we can`t know it as a result of rational analysis. It is not an empirical claim – or, at least, not in any obvious sense of `empirical` - thus its truth is not given in perception. So how do we detect disruptions in the moral equilibrium? An obvious, if slightly unexciting, answer is: *intuitively*. Above, I suggested that moral philosophers are likely to understand (B2) as capturing a foundational moral *intuition*. And when a proposition captures an intuition about something then grasping the truth of that proposition happens *intuitively*. We can do better here though.

In 1963, John Rawls published a paper called `The Sense of Justice`.[182] Nowhere in the paper does he give an explicit definition of a `sense of justice` but the context of the paper makes it quite clear what he takes it to be. At places, Rawls`s usage of the notions of `just/justice` and `fair/fairness` indicates that he takes the notions as more or less synonymous.[183] A sense of justice then seems to be a capacity of persons to detect whether a state of affairs is or isn`t just/fair. Taking *fairness* as synonymous with *moral equilibrium* then gives us a sense of justice being a capacity of persons to detect disruptions in the moral equilibrium.

For our purposes, there are several theoretically exciting things about Rawls`s concept of the sense of justice. One of them is that it refers to a capacity that `may be viewed as the

[179] *Fairness* is `axiomatic` in the sense that its grasp is a necessary precondition of any non-trivial moral theorizing.

[180] Something being *fair* would then translate as *being in moral equilibrium*.

[181] This point doesn`t apply to *unfairness*. Some moral states of affairs can be more *unfair* than others. Fairness, however, doesn`t allow any scope. It is infinitely sensitive to disruption and tends to abruptly turn into unfairness.

[182] John Rawls, `The Sense of Justice`, *The Philosophical Review*, 72, no. 3 (1963), pp.281-305.

[183] See J. Rawls, `The Sense of Justice`, p.282-283.

result of a certain natural development`.[184] Rawls makes use of a plausible theory of child development to show how the sense of justice `arise[s] from our primitive natural attitudes`.[185] Throughout the paper he develops this point concluding, ultimately, that:

> […] a person who lacks a sense of justice is also without certain natural attitudes and certain moral feelings of a particular elementary kind. Put another way, one who lacks a sense of justice lacks certain fundamental attitudes and capacities included under the notion of humanity.[186]

A sense of justice is thus not something of an uncertain ontological status, but a capacity deeply embedded in the psychological constitution of a mature human being. It is, in this sense, an objective capacity, not an invention of a moral theorist. It gets employed in our practice of holding others accountable. And it is, ultimately, this sense of justice we possess that informs us about the truth of (B1), (B2) and many other moral claims.

Another theoretically exciting thing about the sense of justice is that it has a normative force. Not only does it detect disruptions in the moral equilibrium, it commands us to restore them. Rawls doesn`t talk explicitly about normative force in connection with the sense of justice. There are, however, passages in the paper that strongly suggest that he wouldn`t object taking the sense of justice as involving a normative force. The most suggestive is perhaps the passage where he gives an answer to the question of `what accounts for men`s acting on their duty of justice in particular cases`.[187] Rawls answers:

> When they have a sense of justice, an answer is that they accept the principles of justice and regard themselves bound to act in accordance with schemes of co-operation which satisfy these principles when it comes their turn.[188]

The plausibility of the claim that the sense of justice involves a normative force can perhaps be best demonstrated by an appeal to our experiences when witnessing something very unfair playing out in front of our eyes. We become indignant, and we feel that *something must be done* in response to the unfairness of the situation (even though we might not – for various reasons – do anything). The *something-must-be-done* feeling is a common and natural response to many morally significant situations: it is our sense of justice commanding to restore the moral equilibrium.

4.6.3 Holding others accountable is motivated by desert/fairness

[184] Ibid., p.281.
[185] Ibid., p.285.
[186] Ibid., p.299.
[187] Ibid., p.298.
[188] Ibid.

The claim about the sense of justice involving a normative force lies at the core of (B3), i.e. at the core of the claim that: holding others accountable is motivated by *desert/fairness*. The phrase `holding others accountable is motivated by *desert/fairness*` is a compact one. It can be unpacked and paraphrased in the following way: in holding others accountable, our sense of justice commands us to restore the moral equilibrium. I believe the reader will not find the paraphrase objectionable. `Being motivated` (by something) entails a reference to, very broadly speaking, a *force.* `Being commanded` (by something, to do something) entails the same. The notion of the `sense of justice` specifies the *locus* where the *force* originates. `Moral equilibrium` is synonymous with `fairness`. `Holding others accountable` is a notion that refers to a morally significant context. We have established that holding others accountable involves imposition of pain. The practice of holding others accountable essentially involves imposition of pain because that is what – as our sense of justice informs us and commands us to do – restores the moral equilibrium.[189]

Now we have all the pieces ready to recap in a form of a brief argument towards (B3):

i. The upshot of the discussion of *the neurocognitive scenario*: the most plausible explanation of why it *feels right* to impose pain on the wrongdoer is that it is what the wrongdoer *deserves*/what is *fair* to do in the scenario.

ii. At the heart of the *feeling right* mentioned in (i) lies what John Rawls calls a `sense of justice` - a common and essential human capacity to detect disruptions in moral equilibrium (that is, disruptions in *fairness*).

iii. The sense of justice involves a *normative force*, i.e. it *commands* or *motivates* us to act in a certain way. More specifically, it commands or motivates us to restore the moral equilibrium that has been disrupted by a wrongdoing (the moral equilibrium gets restored by, among others, imposing the right amount of pain on the wrongdoer).

Once (i) – (iii) are accepted, we can unpack the phrase `we are motivated by *desert/fairness*` as a phrase saying that `our sense of justice commands us to restore fairness/moral equilibrium`. The two phrases are then to be treated as equivalent. Consequently, the

[189] This last point seems to imply that the wrongdoing in response to which we hold the wrongdoer accountable must essentially be a pain-causing act. It seems clear that the pain imposed in the act of holding accountable will restore the moral equilibrium only if it has been disrupted by a pain-causing wrongdoing. I think this is correct as I believe that any wrongdoing must be essentially a pain-causing event to be a *wrong*doing at all. I could be wrong here but that would still not affect in any way my claim that imposition of pain has a restorative effect on the moral equilibrium. The truth of that claim is given to us intuitively – as shown by our response to the *neurocognitive scenario* - and not as a result of an argument.

argument behind the latter – i.e. steps (i) – (iii) - sustains the former, and we are able to conclude (B3): When holding accountable, we are motivated by *desert/fairness*.

4.7 Conclusion

At the beginning of this chapter, I promised to distinguish between *right* and *wrong* instances of holding responsible and provide, afterwards, an argument towards the claim that it is only the *wrong* instances of holding responsible that require the free will assumption. Let me prepare the ground by a way of the following recap. I set myself the task of establishing whether there is a kind of holding responsible that calls for the free will assumption. I have identified one such kind – *holding accountable* – as in need of the assumption. The path towards the identification revealed that we humans are equipped with the sense of justice which commands us to engage in this particular kind of holding responsible, i.e. in holding accountable. Holding accountable is intrinsically about restoring fairness or moral equilibrium disrupted by a wrongdoing, and the sense of justice commands us to engage in a holding accountable exactly because one of its ultimate functions is to push us towards the restoration. This last point is what the phrase `in holding others accountable we are motivated by *desert/fairness*` is meant to capture. Typically, the sense of justice commands us to impose the amount of pain that is proportionate to the amount of pain generated by the wrongdoing. It commands us to do so because that`s what is *fair*, i.e. what restores the moral equilibrium.

Chapter 5: *Right* and *wrong* holdings responsible, and the free will assumption

5.1 Introduction

The discussion in the previous chapter has yielded (among other things) the conclusion that there is a kind of holding responsible that calls for the free will assumption. It is holding responsible as *holding accountable*. Holding a person accountable is *unfair* if the person lacks free will, that is: the possession of free will by the person held responsible is a necessary condition of the holding to be *fair*.

In this chapter, I will argue that there is another necessary condition of a *fair* holding. That condition, however, cannot be met due to certain constraints intrinsic to the moral nature of man. If I am right that there is such a necessary condition of a *fair* holding accountable that cannot be met, then holding a person accountable can never be *fair* regardless of whether or not the person held accountable possesses free will.

5.2 Disruptors

The domain of fairness or moral equilibrium is sensitive to a variety of *disruptors*. A *disruptor* is anything that our sense of justice identifies as bringing or keeping the moral equilibrium out of balance. Any wrongdoing that imposes pain is such a disruptor, and our sense of justice commands us to bring the equilibrium back in balance by imposing a proportionate degree of pain on the wrongdoer (in the act of holding accountable, or in the act of punishment for more serious wrongdoings). At the same time, imposing a proportionate degree of pain on a wrongdoer who lacks free will, i.e. who couldn`t have avoided doing the wrongdoing, not only would not restore the moral equilibrium, it would represent an additional disruptor. Similarly disruptive would be imposing pain onto a wrongdoer who couldn`t have known that her actions would result in causing pain.

We can make a little distinction here regarding the notion of a disruptor. There seem to be two basic groups of them: disruptors as original wrongdoings and disruptors as wrong (i.e. disruptive) responses to the original wrongdoings. I shall call the former *the target disruptors* and the latter *the targeting disruptors*. The target disruptors are, for instance, all those morally significant actions that we teach our children to avoid, such as stealing or damaging someone else`s stuff or various ways of hurting others, such as beating them, ridiculing or disrespecting them, manipulating them, etc. The targeting disruptors are, for instance and as mentioned above, those *responses* (to the target disruptors) that impose pain on a wrongdoer who lacks free will or was ignorant regarding the harmful effects of her actions.

I shall take it as uncontroversial that there is no principal moral difference between a target disruptor and a targeting one: both disrupt the moral equilibrium; the existence and strength of both gets detected by our sense of justice, both are to be avoided or eliminated if detected. The last point, in other words, is: an instance of holding accountable or punishing that is *unfair* (disrupts the moral equilibrium) is a wrongdoing that must be avoided.[190]

Now, what if there are reasons to believe that any instance of holding accountable or punishing is *always* and in principle *unfair* and, as such, morally wrong? This question, and the worry it raises, is, of course, what motivates the interest in the problem of free will because the impossibility of free will would constitute such a principal reason. If free will is impossible then *whenever* we hold another accountable or punish, we act unfairly, we generate a targeting disruptor: we do what is morally wrong. The implicit hope among the philosophers working on the problem of free will is that once we formulate an account of free will that allows for its existence in the causally determined world, there won`t be any other principal reason to doubt the moral rightness of our practice of holding others accountable.

I am convinced, however, that there are other principal reasons - apart from the issue of the (im)possibility of free will – that render any instance of holding accountable *unfair*. This means that even if a successful account of free will is formulated[191] our practice of holding others accountable will still remain *essentially* unfair. Should I be correct and able to convince you regarding the claim that there indeed are those other principal reasons to believe that the practice is essentially unfair, it would, I believe, remove an important obstacle towards accepting my account of freedom in agency. My account doesn`t help with grounding the practice but the practice is ungroundable anyway thus my account shouldn`t be dismissed because of this particular theoretical impotence. I proceed to the convincing.

5.3 Blame, holding accountable, and the moral standing

There has recently been increased interest in the topic of *blame* among moral theorists.[192] The issues discussed in connection with *blame* are the nature of blame, its function and the conditions for the appropriateness of blaming.

Regarding the appropriateness of blaming others for their wrongdoing, there seems to be a consensus about what general facts have to be taken into consideration when assessing the appropriateness. The facts relevant for the appropriateness assessment can be sorted into

[190] And, of course, as such, it calls for a restorative action of some sort: a holding accountable or a punishment.

[191] That is, an account that will morally ground the practice of holding accountable in a causally determined world.

[192] For an excellent collection of essays on blame see D. Justin Coates and Neal A. Tognazzini (eds), *Blame: Its Nature and Norms* (New York: Oxford University Press, 2013).

roughly three groups: (a) facts about the blamer, (b) facts about the blaming interaction, and (c) facts about the person being blamed. In what follows, I will make use of a concept that plays a crucial role in discussions of the first group, i.e. in discussions of facts about the blamer. The concept I have in mind here is the concept of *moral standing*.

There are plenty of real and imaginary scenarios in which even if the agent was blameworthy, and even if all procedural norms were followed, it would be morally inappropriate for some people to blame. For blaming to be appropriate, not only the target person has to be *blameworthy* but the blamer herself has to be what Marilyn Friedman calls *blame(r)worthy*.[193] For instance, a serial thief blaming another for a petty theft would strike us as such a case of morally inappropriate blaming exactly because the serial thief is clearly not *blame(r)worthy*. `Who are *you* to criticize another for *that*?`, we would want to interject if we witnessed an instance of such a morally inappropriate blaming. What gets brought into focus and questioned in such an interjection is the *moral standing* of the blamer. The background intuition here is that if the moral standing of a person is in some relevant sense compromised then it would be morally inappropriate for that person to blame another.

Uncontroversially, holding accountable involves blaming. Thus, to the extent to which the appropriateness of blaming depends on the moral standing of the blamer, holding accountable too will depend – with respect to its appropriateness – on the moral standing of the holder. If the moral standing of a holder has been compromised, then the holding is (morally) inappropriate or wrong. A husband involved in a long lasting extra marital affair attempting to hold his wife accountable for a one-night stand with her ex will strike us as an instance of a morally inappropriate holding accountable because of the compromised moral standing of the husband. Such an instance of holding accountable is a targeting disruptor of the moral equilibrium: it is wrong and should be avoided.

5.4 Moral standing of human agents as essentially compromised

The moral appropriateness or rightness of holding another accountable depends crucially on the moral standing of the holder. Now, what if the moral standing is *always* compromised? What if fundamentally compromised moral standing is in some sense *essential* to being a human? If this was the case, surely, we would have to conclude that our practice of holding others accountable is morally, and irremediably, wrong. There are reasons to believe that our moral standing is fundamentally compromised in such a way. Below, I shall formulate an argument to this effect. The argument will employ (a version of) a thesis known as The Principle of Plenitude and a thought experiment that allows its applicability in the context of

[193] Marylin Friedman, `How to Blame People Responsibly`, *The Journal of Value Inquiry*,47, no. 3 (2013), pp.271-284 (p.272).

evaluating the moral standing of man. Let me first give you a rough outline of the argument so that the logical role and the mutual relation of its individual steps is clear before they are later discussed in detail.

An outline of the argument:

1. The Principle of Plenitude (PP): For some states of affairs *s*, if *s* is possible then there is a time at which *s* obtains.
2. A state of affairs that has obtained is a possible state of affairs.
3. Plausibly, if a state of affairs involving a member *m* of a kind *k* is a possible state of affairs then a relevantly similar state of affairs involving any other member of the kind *k* is possible too.
4. Some men did or have done things that (have) corrupted their moral standing.
5. It is possible for any man to have their moral standing corrupted. [from 2, 3, 4]
6. [*A thought experiment designed to show that*] One`s (present) moral standing gets corrupted (not only by the past and present wrongdoings but even) by one`s future wrongdoings.
7. [*For reasons discussed below*] (PP) can be applied to (5), that is: the status of one`s moral standing belongs to the states of affairs that (PP) applies to.
8. For all men there is a time at which they do things that corrupt their moral standing. [from 1, 5, 7]

Therefore

9. The moral standing of man is essentially corrupted. [from 6, 8]

The individual steps of the argument will be discussed in dedicated subsections below.

5.4.1 The Principle of Plenitude

The Principle of Plenitude as I will understand it here is grounded in what is sometimes called the `statistical interpretation of modality`.[194] The statistical interpretation (or model) of modality can be spelled out roughly in the following way: what is necessary is always actual, what is impossible is never actual and what is possible is at least sometimes actual. The Principle of Plenitude is then a thesis about a certain kind of relation between *possibility* and *actuality*. A good first approximation of the principle is given by Hintikka:

[194] The term `statistical interpretation of modality` was introduced into modern discussion of modality by Oscar Becker. See his *Untersuchungen über den Modalkalkül* (Meisenheim am Glan: Anton Hain, 1952)

> [*A*]*ll genuine possibilities*, or at least all possibilities of some central and important kind, *are actualized in time*. Any such possibility thus has been, is, or will be realized; it cannot remain unrealized through an infinite stretch of time; in a sense, everything possible will happen in the long run.[195]

A version of the principle (and/or the modal intuition behind it) has been endorsed by philosophers such as Plato, Aristotle, Epicurus, Augustin of Hippo, St Anselm, Thomas Aquinas, Giordano Bruno, Spinoza, Hobbes, Leibniz, Kant and Russell.[196] Kant, for instance, seems to endorse the principle in the following passage:

> The Schema of possibility […] is the determination of the representation of a thing at any time whatsoever. The schema of reality is the existence at a given time. The schema of necessity is the existence of an object at all times.[197]

And Russell, in his characteristically clear and unambiguous style, asserts:

> One may call a propositional function *necessary*, when it is always true; *possible*, when it is sometimes true; *impossible*, when it is never true.[198]

And on the next page in the same paper, Russell casually equates the notion of *sometimes* with that of *possibility*:

> It will be out of this notion of *sometimes*, which is the same as the notion of *possible*, that we get the notion of existence.[199]

Obviously, the first approximation of the principle given above allows different versions depending on what kind of possibility one has in mind. Possible states of affairs? Possible kinds? Possible particulars? The plausibility of the principle might depend on the kind of

[195] Jaakko Hintikka, `Aristotle on the Realization of Possibilities in Time` in Simo Knuuttila (ed.), *Reforging The Great Chain of Being* (Dordrecht: Springer-Science+Business Media,B.V., 1981), pp 57-72, (p.58).

[196] All the names listed above, apart from that of Aristotle, Hobbes and Russell, are the names given by the historian of ideas Arthur Lovejoy in his *The Great Chain of Being* (Harvard University Press, 1936). Lovejoy wasn`t a philosopher and the list is somewhat controversial. Thus, for instance, Jaakko Hintikka suggests that it wouldn`t be correct to take Plato as an adherent of the principle. Regarding Kant, Hintikka argues that Kant endorses the principle only in his pre-critical writings. (I tend to disagree with Hintikka on this point as there is, in my view, a passage in The Critique where Kant seems to say something which is very close to the Principle. I quote the passage below). And as for Leibniz, although there is a version of the principle that Leibniz embraced, it is a version that is rather different form the one given above. In fact, Leibniz explicitly rejected the version that we shall work with here. For details on these points see Jaakko Hintikka, `Kant on "The Great Chain of Being" or the Eventual Realization of All Possibilities: A Comparative Study`, *Philosophic Exchange*, 7, no.1 (1976), pp. 69-89.

[197] Immanuel Kant, *Critigue of Pure Reason*, F. Max Müller (trans.) (Doubleday, Garden City, N.Y.: 1966) p.125.

[198] Bertrand Russell, `The Philosophy of Logical Atomism` in Bertrand Russell, *Logic and Knowledge. Essays 1901-1950*, ed. R.C.Marsh (Allen&Unwin, London: 1956), p.231.

[199] Ibid., pp. 231-232.

possibility that it is applied to. Jonathan Barnes, for instance, thinks that the principle doesn`t apply to states of affairs involving perishable particulars because out of the numerous possible ways that a perishable particular can perish only one can happen: once the particular has perished, all the remaining (and previously) possible ways of perishing become impossible regarding that particular because the particular is no more.[200] Later in this chapter (when discussing step 7) I will address this issue as I need the principle to apply to individual humans, that is, to perishable particulars. At this point, however, I have a bigger problem to deal with. The problem is that in the contemporary analytic literature, the statistical understanding of modality has been fully replaced by interpreting modality in terms of what is generally known as possible worlds semantics – and modality interpreted in terms of possible world semantics doesn`t involve any temporal references. Consequently, there is virtually no discussion of the principle of plenitude to be found among contemporary analytic philosophers writing on modality.[201] Thus, not only am I unable to simply appeal to the principle and proceed to discussing the following steps of the argument, I must assume that the principle will be perceived by my reader as weird at best and totally implausible at worst. A way forward at this point is to provide an argument for the principle. One such argument can be extracted from Jonathan Barnes` comments on Jaakko Hintikka`s *Time and Necessity: Studies in Aristotle`s Theory of Modality*.[202]

Barnes gives his argument (or what could be reconstructed as his argument from the review) for the principle in the context of discussing Hintikka`s interpretation of some of the aspects of Aristotle`s treatment of modality. In his review of Hintikka`s paper, Barnes is not primarily concerned with providing an argument for the principle. What he is concerned with is complementing Hintikka`s interpretation of certain passages from Aristotle in which the Stagirite seems to argue for the principle with some charitable reading of those passages. In the passages Aristotle`s argument seems rather obscure and this is where Barnes steps in offering a sympathetic reconstruction of Aristotle`s thinking behind the argument.

The argument for the principle as given by Barnes relies crucially on the assumption that, roughly, *that which always obtains is necessary*, or:

A: if *s* always obtains then *s* is necessary

[200] See Jonathan Barnes (review of) J. Hintikka, `Time and Necessity: Studies in Aristotle`s Theory of Modality`, *The Journal of Hellenic Studies*, 97 (1977), pp. 183-186, (p.184).

[201] Even though there are historians of philosophical ideas and their development who discuss the principle in an analytic way. One of them is the above-mentioned Jaakko Hintikka. Another one is, for instance, Jonathan Barnes, whose brilliant argument for the principle will be discussed below.

[202] See Jonathan Barnes` review of J. Hintikka, *Time and Necessity: Studies in Aristotle`s Theory of Modality*, in *The Journal of Hellenic Studies*, Vol. 97 (1977), pp. 183-186.

The assumption is, however, rather controversial. It will be pointed out that it is easily conceivable that *s* always happens and yet *s* is contingent. Moreover, the way modality is, in (A), tied to temporality seems to be uncannily similar to the way in which we wish to tie modality to temporality in (PP). Thus, (A) could be seen as begging the question. Clearly, an argument for (A) is needed, and Barnes gives us one. Unfortunately, there seem to be several problems with the argument. The most serious one is the fact that the argument depends crucially on Aristotle`s definition of possibility – a definition that is rather implausible. Aristotle thinks, roughly, that something is (can be defined as) possible if and only if nothing impossible results from its actualization; Barnes puts this more rigorously in the following way:

> T: *s* is possible =(df) there is no state *s** such that *s** is impossible and if *s* obtains then *s** obtains[203]

It is immediately obvious, I believe, that (T) doesn`t work as a *definition* of possibility, although it perhaps works as a *necessary condition* of possibility. Now, as mentioned above, Barnes`s argument for (A) turns on the truth of (T). As a *definition* though, (T) is false. This collapses the whole argument for (A) as formulated by Barnes.[204] A different argument is needed.

I have thought hard about (A) and I have come to suspect that there is no good argument showing how something`s necessity could be *conceptually* derived from that something always obtaining. This, however, doesn`t mean that (A) cannot be argued for. An alternative and common way of arguing for a claim is to show that accepting its falsity is theoretically too costly.

So, what are the theoretical costs of denying (A)? One way of denying (A) is this:

> A*: *s* always obtains and (yet) *s* is not necessary

First, notice that (A*) is a claim that will be embraced by a Humean. A Humean believes that the world has no nomological structure, that is, she doesn`t believe that there is any law-likeness out there in the world. She sees only contiguity, temporal priority and constant conjunctions where others see Laws of Nature being instantiated. Uncontroversially, the notions of a `*nomological* structure` and the `Laws of Nature` entail an appeal to the modal property of necessity. Thus, a Humean`s denial of the world having a nomological structure

[203] Ibid., p.185.

[204] Barnes is, of course, aware of the implausibility of (T). The task he sets himself in the review is to reconstruct what he thought was Aristotle`s argument. Barnes`s faithful reconstruction inherits the problematic definition of possibility from Aristotle.

or of (being governed by) the Laws of Nature entails her denying that there is any necessity out there in the world.[205] Consequently, she will endorse (A*).

The Humean view according to which there is no nomological structure (or Laws of Nature) behind the constant conjunctions that we observe is called a *regularity theory*. A regularity theory holds that there is nothing *beyond* the regularities that somehow holds the world together: nothing that *underlies* them, nothing that *explains* them. Now, it should be noticed that a regularity theory is a rather extreme theory. It implies that *s*`s regular and exceptionless obtaining is not grounded in the necessity that is intrinsic to the Laws of Nature but - given an infinite time - due to a mindbogglingly extraordinary coincidence. Galen Strawson points out this implication here:

> According to [regularity theories]…, the regularity of the world`s behaviour is, in a clear sense, a complete and continuous fluke. It`s not just that we don`t know whether or not there is any reason for it in the nature of things. According to [regularity theories], there is definitely no reason for it in the nature of things.[206]

Strawson has no patience with this view calling it `utterly implausible` and `absurd`:

> [T]he theory is utterly implausible in asserting categorically that there is no reason in the nature of things for the regularity of the world … it is absurd to say – to insists – that there is definitely no reason in the nature of things why regularity rather than chaos … occurs from moment to moment.[207]

I am in total agreement with Strawson here. There is more to be noticed about the view though. Regularity theory seriously clashes with one of the foundational assumptions behind any theorizing about the world: the assumption that the world can be made sense of; that it is *explicable*. In such a world, coincidences of this magnitude cannot exist. The assumed explicability of the world commits us to assuming that the world has a nomological structure, appeals to which play crucial role in anything that counts as an explanation of the world. What is at stake here is not only the project of natural sciences to understand and explain the world, but also the prospects of *philosophizing* about it. A Humean – that is, a regularity theorist – rejects any talk of nomological structures or Laws of Nature because she denies the reality of causation. Without causation, there is no real *connection* between things that would allow the transfer of necessity between them. Thus, without causation, there are no *Laws* of Nature and nothing *nomological* about the world. But one cannot philosophize without taking causation seriously. As Helen Beebee – who is a regularity theorist – admits,

[205] This is not to say that a Humean denies *conceptual* necessity. She might do so, or she might not, depending on her other philosophical commitments. I focus on the *empirical* necessity here as I am arguing towards a modal claim about humans, that is, about empirical entities.
[206] Gallen Strawson, *The Secret Connexion: Causation, Realism, and David Hume* (Oxford: Clarendon Press, 1989) p.21.
[207] Ibid., pp.21-22.

> [there is a] huge range of fruitful philosophical theories that *do* appeal to causation: we have causal theories of perception, reference, action and knowledge; functionalist theories of the mind, consequentialism; and so on, and on.[208]

A Humean is a philosopher. Thus, endorsing the regularity theory will cost her dearly: Not only will she have to embrace a rather unattractive claim that natural sciences don`t *explain* but merely *describe* the world, she will also have to ditch (on pain of being inconsistent) a `huge range of fruitful philosophical theories`, and, perhaps, even stop philosophizing altogether because, as Beebee notices, `trying to do philosophy without ever using the concept of causation is practically impossible`.[209]

Now, it is somewhat surprising that Strawson`s objections, given how damaging they are, haven`t elicited much response from regularity theorists. A notable exception is the response given by Helen Beebee in the paper I have quoted from above. The paper accurately and fairly presents Strawson`s objections to regularity theories and raises several important and correct points about the objections that a Strawsonian should take into account. That`s not all. Beebee takes on what I think is Strawson`s strongest objection – an objection that plays a key role in my argument for (A). Above, I have appealed to the extreme implausibility and unpalatability of the implication (of regularity theories) that the highly complex orderliness of the world is a result of a mindbogglingly extraordinary luck. Strawson invites the reader to appreciate the absurdity of the implication through considering the following analogy:

> [Imagine that] a true randomizing device determines the colour value of each pixel on a standard 800 x 400 computer screen, running on a ten-times-a-second cycle – so that each pixel can take any colour value for each 1/10th second period. On the screen it appears that there is a film showing. A woman enters a house, walks over to a stove, and puts on a kettle. Life – a world, as it were – goes on in an ordered, regular fashion, exactly as regularly as in our own world. But the image is being generated by the true randomizing device. It is pure fluke that what happens on the screen appears to tell a coherent story of a regular, ordered world, rather then filling up with – or suddenly switching to – a fizz of points of colour.[210]

The analogy is powerful, vividly exposing the theoretical costs of endorsing a regularity theory. Beebee`s response is smart. She doesn`t attack the analogy itself - in fact, she urges a regularity theorist that she `must […] accept that from a metaphysical point of view the analogy is a pertinent one`[211] - she, instead, argues that the implication can be *tolerated*. Her argument in this respect is, roughly, this: True, a regularity theory comes at the cost of

[208] Helen Beebee, `Does Anything Hold the Universe Together?`, *Synthese*, 149, no. 3, *Metaphysics in Science* (Apr., 2006), pp.509-533, (p.510).

[209] Ibid., p.510.

[210] Strawson, *The Secret Connexion: Causation, Realism, and David Hume*, p.24.

[211] Beebee, `Does Anything Hold the Universe Together?`, p.527.

accepting that the highly complex orderliness of the world is due to just a massive – and ongoing – luck. However, that`s nothing to be much upset about because we have already learned how to tolerate `outrageous runs of luck`.[212] Consider your own life. You are alive as a result of an extremely long series of lucky events. Think of all those things that had to happen in order for you to be born. On countless occasions, your parents might have done something that would have prevented you having been conceived, or they might have not even met in the first place. The same goes for your grandparents on whose actions the existence of your parents – thus yours too - depends. Ultimately, your existence and everyone else`s existence depends on that spectacularly lucky streak of events that resulted in Earth being a place that supports life. Now, when you start thinking about all this, how much does it really bother you? Most likely, not much at all. You don`t really think there is, or must be, any `reason` why things happened in a way that ultimately led to your coming into existence.

This is an intriguing reply even though I don`t think it works. Let`s have a closer look at what is going on here. Strawson formulates a thought experiment designed to expose the extreme implausibility of the claim that a purely random process can, at the same time, be a process that is highly ordered and keeps being so for very long periods of time. In response, Beebee invites the reader to consider their own life to see that it, too, despite being ordered and coherent, is a result of a long series of lucky events. She notices that it doesn`t seem to bother us much that luck plays such a fundamental role in our lives and concludes that we, in fact, already know how to tolerate the seemingly intolerable implication of regularity theories. Beebee`s reply turns on being presented as an *analogy* to Strawson`s thought experiment. And if it is an analogy then whatever the analogy shows can be – by analogy – said to be shown about the thought experiment. In our case: if we stay unperturbed about our lives being a massive fluke (as shown by the analogy), then why be perturbed about (a) a movie with a coherent story being generated by a randomizing device, and (b) about, ultimately, the implication of regularity theories? The problem with the reply is that the little consideration that Beebee offers as an analogy to the thought experiment is an analogy only seemingly.

The intuitive force of Strawson`s thought experiment depends crucially on contrasting a *true randomizing device* with the high level of orderliness and coherence of a movie. Beebee, however, doesn`t mention any randomizing device at all. She, instead, talks about *luck*, and contrasts it with the orderliness and coherence of one`s life. Presumably, a *true randomizing device* and *luck* are treated as conceptual analogues here. *Luck*, however, is a notion that is ambiguous in a way that a *true randomizing device* isn`t – it has both a metaphysical and an epistemological reading. *True randomizing device*, on the other hand,

[212] Strawson, *The Secret Connexion: Causation, Realism, and David Hume*, p.26.

has (at least in the context of the thought experiment) only a metaphysical reading. Now, it certainly feels perfectly natural to *see* your own life as a *miracle* of sorts: so many things could have gone wrong over such a long time, and if they had, you wouldn`t have been born. But they didn`t, and once you stop for a moment to appreciate this, you can`t but feel extraordinarily *lucky*. This feeling, however, is just due to you having an epistemological access to only a tiny fragment of the total facts that, as a whole, *produced* you. If you knew the totality of those facts you would feel about as lucky as you feel when you clap your hands and it produces a sound, that is: you wouldn`t find anything lucky about it. The huge gaps in your knowledge about the past events that gave rise to your existence give a strong impression that those relatively very few facts you know are somehow floating in the ocean of randomness. This ocean of randomness, however, is, in fact, just your inevitable ignorance regarding the totality of events that, ultimately, produced you. If you knew the totality of the facts, the ocean would evaporate. Beebee`s reply by way of offering and exploiting an analogy to Strawson`s thought experiment doesn`t work because Strawson contrasts a highly ordered and coherent state of affairs with *metaphysical* luck, while Beebee contrasts it with (what I call here) *epistemological* luck.

In little bit more detail, the problem is this. Arguably, most people know, often perhaps in some pre-conceptual way, that the kind of *luck* they accept as being involved in their own and others` lives is something like the kind of luck I have qualified above as *epistemological* luck. Or, *at least*, they would resist understanding the luck involved in their lives as being conceptually equivalent to a *true randomizing device*. Why do I claim this? It is a safe bet to expect that the vast majority of people would find Strawson`s thought experiment convincing.[213] That is, they would agree that highly ordered and coherent states of affairs lasting for long periods of time cannot emerge out of a *truly random* process. They, at the same time, accept that they are very lucky regarding their lives, which can only mean that people normally don`t understand *luck* as a *truly random* process.[214] Thus Beebee cannot claim that we already know how to tolerate luck in our lives in the sense of tolerating the thought that our lives emerge out of a truly random process. She can claim so only in terms of (what I call) epistemological luck. In this sense, however, her reply misses the target.

[213] Recall that Helen Beebee herself agrees that the thought experiment `is a pertinent one`.

[214] The following uncharitable answer is possible: The majority of people have inconsistent intuitions. Therefore, they will be convinced by Strawson`s thought experiment and yet see their lives as lucky in the sense of being truly random. This reply is rather unattractive. Accepting that most people have inconsistent intuitions severely undermines any appeal to intuitions in philosophical arguments. Some of the most important arguments in philosophy rely on an appeal to (rational/conceptual) intuitions. Thus, the uncharitable reply would be far too costly.

To the best of my knowledge, Beebee`s response to Strawson`s thought experiment is the strongest challenge to it that can be found in philosophical literature. It fails nonetheless. Consequently, a Humean objection to (A) fails too.

Once the truth of (A) has been established, the rest of the argument for (PP) is relatively straightforward. Recall:

> A: if *s* always obtains then *s* is necessary

> The Principle of Plenitude (PP): For some states of affairs *s*, if *s* is possible then there is a time at which *s* obtains.

(PP) is then readily derivable from (A) in the following way. Suppose that *s* is a contingent state of affairs. That is, suppose that *s* is a possible but not necessary state of affairs. Then *s*`s not obtaining is a contingent (that is, possible but not necessary) state of affairs. If that is so, then *s*`s not obtaining is not necessary. And from this it follows, by (A), that `*s*'s not obtaining does not always obtain; hence there is a time at which the non-obtaining of *s*'s not obtaining obtains; i.e. there is a time at which *s* obtains`.[215] This might feel too condensed, so let me unpack it here a little:

i. Suppose: *s* is (a) contingent (state of affairs).
ii. (i) entails that *s*`s not obtaining is contingent.
iii. (ii) entails that *s*`s not obtaining is not necessary.
iv. A: if *s* always obtains then *s* is necessary .
v. By (A): if *s*`s not obtaining is not necessary, then *s*`s not obtaining doesn`t always obtain.
vi. If *s*`s not obtaining doesn`t always obtain, then there is a time at which *s*`s not obtaining doesn`t obtain.
vii. To say that `there is a time at which *s*`s not obtaining doesn`t obtain` is to say that `there is a time at which *s* obtains`.

I take it that it has been proved that (for some states of affairs)[216]: if *s* is (a) contingent (state of affairs) then there is a time at which *s* obtains. This is, for our purposes here, close enough to (PP), thus I shall, from now on, assume the truth of (PP).

[215] Ibid., p.185.

[216] A reminder: (PP) doesn`t apply to all kinds of states of affairs. As mentioned above, (PP) won`t work for states of affairs involving, for instance, perishable particulars. This deficiency will be addressed when discussing step 6 below.

Before I proceed to the discussion of step (2) regarding the argument behind the claim that the moral standing of man is essentially corrupted, let me briefly address a little issue regarding the plausibility of (PP). I strongly suspect that many readers – even those that have been convinced about the validity of the argument - will find (PP) and its implications just too fantastic to swallow. As Barnes notes:

> According to (PP), elephants will tell each other human jokes, the first daffodils of autumn will appear when the leaves fall upwards to the trees, and pigeons will hunt cats through city backyards.[217]

This, indeed, is a rather unpalatable corollary to (PP). Is it, however, unavoidable? It isn`t, I believe. (PP) is a thesis about (some) possible states of affairs, i.e. it tells us something interesting (and perhaps unexpected) about (some) possible states of affairs. It doesn`t, however, come with any prior commitment to what states of affairs count as *possible*. In this respect it is entirely up to the reader to decide what states of affairs she accepts as possible. In other words, the reader can, if she wishes, avoid the above mentioned unpalatable corollary to (PP) by refusing to accept as possible states of affairs the ones in which elephants tell each other human jokes, the leaves fall upwards to the tress, and pigeons hunt cats through cities. I, for one, am far from sure that these count as possible states of affairs. A deeper point here is this. (PP) could be taken as a metaphysical definition of possibility – this is, I believe, just a matter of choice. Once accepted as a metaphysical definition of possibility, it pre-empts any objection that appeals to the implausibility of its implications of the kind Barnes gives because the logic of such a definition serves as a constraint on what counts as a possibility in the first place. Be that as it may, the kind of possibility we will apply (PP) to below is nowhere close to as fantastic as the possibilities conceived of by Barnes.

5.4.2 A state of affairs that has obtained is a possible state of affairs

There is an obvious sense in which this assertion is correct. If *s*`s not obtaining is necessary, then *s* never obtains. Hence, if *s* has obtained then *s*`s not obtaining is not necessary. And, if *s*`s not obtaining is not necessary then *s* is possible. This little argument appeals to a straightforward logical relation between the concepts of `something never obtaining`, `something being necessary` and `something being possible`. Now, things get a bit complicated once we go beyond a merely conceptual reading of the assertion. One might wonder whether and in what sense the assertion *works* out there in the world. Does the assertion as it stands imply that, for instance, if a state of affairs has obtained, then it is

[217] Jonathan Barnes (review of) J. Hintikka, `Time and Necessity: Studies in Aristotle`s Theory of Modality`, *The Journal of Hellenic Studies*, 97 (1977), pp.183-186, (p.184).

possible that it will obtain again? Surely, one would be justified in reading it in this way. A moment`s reflection reveals, however, that, at least for some states of affairs, this cannot be true. There are possible states of affairs that, (a), involve perishable particulars and, (b), involve events that cause the involved particulars to perish. Clearly, once such a possible state of affairs has obtained and the particulars involved have perished, it is impossible for that state of affairs to obtain again (that is, to obtain in future). Thus, there are states of affairs that have happened and yet are *in a sense* impossible. For reasons that will become obvious later, I need to be able to read the assertion as implying that if a state of affairs has obtained then it is possible it will obtain again. To allow that reading, the assertion needs to be qualified in something like the following way:

> AQ: If a state of affairs has obtained and if its relevant subject(s) has/have survived the obtaining, then it is possible that the state of affairs will happen again.

Not much turns on the notion of a relevant subject (of a state of affairs), and I don`t intend to give a definition of it here. Its function in the qualified assertion is just to block the application of the assertion onto the states of affairs that happened but cannot possibly happen again because its relevant subject(s) has/have perished. Clearly, the death of Bertrand Russell implies that a state of affairs in which Bertrand Russell has died is, *conceptually* speaking, a possible state of affairs. It is not, however, a possible state of affairs in the sense of (AQ) as this state of affairs cannot happen again simply because the subject of this state of affairs doesn`t exist anymore.

I believe that our ordinary intuitions about possibility are governed by, among others, something like (AQ). Consider the case of climbing Mount Everest without the use of supplemental oxygen. For a long time, it had been hotly disputed whether this was possible at all. Then, on 8 May 1978, Messner and Habeler reached the summit of Mount Everest without the aid of supplemental oxygen. This achievement has established that it is possible for man to climb Mount Everest without supplemental oxygen.[218] It has been established both *conceptually* and in the sense of (AQ). In the sense of (AQ), it has been established because it has happened and the subject of this kind of state of affairs – man – is still around to possibly repeat the performance.

5.4.3 Plausibly, if a state of affairs involving a member *m* of a kind *k* is a possible state of affairs then a relevantly similar state of affairs involving any other member of the kind *k* is possible too

[218] Here I presume as unproblematic to generalize from a member of a species to species as a whole. It sounds very natural to say that a bristlecone pine tree can live (i.e. it is *possible* for it to live) for more than 5000 years on the grounds that just one particular member of this species, Methuselah in California`s White Mountains, has lived that long, even though vast majority of bristlecone pine trees had lived nowhere close that long. I will say more about this in the following section.

I have assumed this in the last paragraph of the previous section where I treated a particular achievement of two Italian mountaineers as indicative of what is possible for man as a species. In a footnote related to that paragraph (footnote 218), I gave a brief consideration in support of this treatment.

Let me repeat and slightly expand the supporting consideration. There is a particular tree in California`s White Mountains that has been named, quite tellingly, *Methuselah*. The tree belongs to the species of a bristlecone pine tree and is believed to be almost 5000 years old. This makes it the oldest non-clonal tree in the world. Now, as far as we know, no other bristlecone pine tree is as old as Methuselah, and the vast majority of the other bristlecone pine trees had lived nowhere close to 5000 years. Yet, it is fairly common to generalize from what we know about Methuselah to what we take as possible about the species that Methuselah belongs to. When you start reading about these amazing trees, you will often come across something like the following perfectly natural sounding statement: `The bristlecone pine can live 5000 years, making it the oldest individually growing organism on the planet, […].[219] This statement is a good example of precisely the kind of a generalization from a single (past) achievement of a member of a species to what is possible for the species as a whole that I endorse here as plausible. The context of the article this statement is taken from makes it clear that `the bristlecone pine` refers to a species, `can` refers to possibility, and the figure of 5000 years relates to Methuselah.

5.4.4 Some men did or have done things that (have) corrupted their moral standing

This is an uncontroversial empirical fact. The list of serious wrongdoings done by millions of men throughout the history is disturbingly long. In a suitable context, any of those serious wrongdoings would be deemed sufficient to critically undermine one`s moral standing.

5.4.5 It is possible for any man to have their moral standing corrupted

This is step (5) of the argument, and it follows unproblematically from the previous three steps. Step (4), when slightly reformulated, says that some members of the species of homo sapiens – that is, some men – have done things that have corrupted their moral standing. Step (2) says that a state of affairs that has obtained is a possible state of affairs. This then entails that a state of affairs in which a member of the species homo sapiens does things that corrupt his/her moral standing is a possible state of affairs. Step (3) allows generalizing from a possibility about an individual member of a species to a possibility about the species as a

[219] Scott Smith, `Scientists: The Future of oldest tree species on Earth in peril`, *AP News*, September 14, 2017 <https://apnews.com/776e453d15674f1e9eb20af289d6e46e/Scientists:-Future-of-oldest-tree-species-on-Earth-in-peril> [accessed 26 June 2020]

whole. Hence, we can conclude that it is possible for any man to have their moral standing corrupted.

5.4.6 One`s (present) moral standing gets corrupted (not only by the past and present wrongdoings but even) by one`s future wrongdoings

On the face of it, this sounds rather unintuitive. How can one`s present moral standing get corrupted by a future wrongdoing, i.e. by a wrongdoing that hasn`t happened yet? The unintuitivness of the claim has to do with our common understanding of how causality and the time arrow relate to each other. Normally, we won`t take future events as being causally efficacious in the present because it seems a fundamental fact that the future is due to the present and not vice versa.[220] And it might seem like that`s what we are being asked to do here: we are invited to consider and accept that one`s *future* wrongdoing affects one`s *present* moral standing. The appearance is misleading though. The relation between one`s wrongdoing and one`s moral standing is not a causal relation, or, at least, not a straightforwardly causal one. I don`t wish to go into the metaphysics of causality here to illuminate the point. Instead, consider the following simple analogy:

You have acquired insider information that in two months the government will introduce a drastic currency reform that will depreciate the value of pound ten times. In response to this information you withdraw all the savings from your bank account and buy gold with it.

Now, there is a sense in which this little story could be described as a future event affecting your present state or actions: a currency reform happening in two months causes you to adopt (now) certain financial measures in response.[221] Thus, as we can see, there is a perfectly natural way of taking future events as affecting the presence.

The analogy, however, will take us only so far. The target claim is that one`s future wrongdoing corrupts one`s present moral standing. In the case of the currency reform, its happening in the future – although it, in a sense, affected my *actions* in the present – hasn`t

[220] This is not to say that this seemingly uncontroversial fact hasn`t been challenged. There is some intriguing literature on the issue of backward causation that seriously discusses the possibility of cases where the effect temporally, but not causally, precedes its cause. See for instance Michael Dummett, "Can an Effect Precede its Cause", *Proceedings of the Aristotelian Society*, 28 (Supplement) (1954) pp.27–44., or Jan Faye, "Causation, Reversibility, and the Direction of Time", in Jan Faye, Uwe Scheffler and Max Urchs (eds.), *Perspectives on Time* (Boston Studies in the Philosophy of Science, Vol. 189), (Dordrecht: Kluwer Academic Publisher, 1997) pp.237–266.

[221] This, of course, is neither the only nor the most natural reading of the little story. It could be insisted that it is rather a present *belief* about a future state/event then the future state/event itself what causally affects your actions in the story. And I wouldn`t want to object to this. My point here is just that there is quite a natural way of taking future states/events as affecting the present, and that this little story is, by way of analogy, a first approximation towards understanding how one`s *future* wrongdoing could be taken as corruptive of one`s *present* moral standing.

done anything to the value of the *currency* in the present. Apart from a few insiders in the government and the Central Bank, no one knows about the planned reform, and, therefore, there are no bank runs that would depreciate the currency. Now, clearly, the currency and the moral standing are counterparts in the analogy. Thus, if we cannot conclude that the currency has devaluated then we cannot, by analogy, conclude that the moral standing has become corrupted. More needs to be done. Consider the following thought experiment:

There is a time-machine device that makes it possible to find out what (if any) moral wrongdoing a person does in future. John`s wife Mary has found out about John`s extramarital affair. Mary reproaches John for cheating on her: `You are a despicable person. How could you betray me like this?! You make me sick.` In response, John turns on the time-machine device. The device informs them that in 4 years from now Mary will also cheat on her husband (although, alas, not John but someone else will be Mary`s husband by then). John breaks the silence: `Who are you to call me a despicable person? You are no different!`.

Now, I believe it will be agreed that John`s reply to Mary`s reproach, and his questioning of her moral standing, is totally appropriate. And if that is so, then we can conclude that one`s future wrongdoing corrupts one`s (present) moral standing.

5.4.7 (PP) can be applied to (5), that is: the status of one`s moral standing belongs to the states of affairs that (PP) applies to

In section 5.4.2 we have identified an ambiguity in the notion of possibility. *Conceptually* speaking, a state of affairs that has obtained is a possible state of affairs. However, the same state of affairs will not be possible in the sense of it possibly obtaining again if it, (a), involves perishable particulars and, (b), the state of affairs is such that when it has obtained, the relevant particulars perished.

Now, in the following section, I will want to conclude that for all men there is a time at which their moral standing gets corrupted. We already know that it is possible for any man to have their moral standing corrupted. At the same time, many (perhaps most) men haven`t, yet, had their moral standing corrupted. This means that if we want to conclude that for all men there is a time at which their moral standing gets corrupted, then that time must be in the future. There is, however, something else waiting in future for all men – their death. Any man is a perishable particular and it is certainly possible that they will perish before they manage to corrupt their moral standing. Surely, it would be extremely implausible to claim that none of the people presently alive will die before they manage to corrupt their moral standing. This possibility represents a serious challenge to my argument, because if it is possible for a man to perish before they corrupt their moral standing then I won`t be able to conclude that for *all* men there is a time (in future) at which their moral standing gets corrupted.

One way to respond to this challenge is to appeal to certain implications that can be extracted from the following assertion:

> L: *Luck* cannot make a difference in one`s moral standing.

Thomas Nagel refers to something like (L) when he says that it is `intuitively plausible that people cannot be morally assessed for what is not their fault, or for what is due to factors beyond their control`.[222] Nagel claims this in the context of discussing what has been known as the Problem of Moral Luck (PML). (PML), roughly, is a problem constituted by an obvious tension between the intuitively compelling (L) and the fact that in our common practice of holding others morally responsible, luck does seem to make a difference (for instance, a drunk driver that ran over a pedestrian will be blamed more than a drunk driver who was lucky that there were no pedestrians around when he was driving home from pub). We don`t need to go into the intriguing details of (PML) here. For our purposes, it suffices to notice that (PML) is a problem taken seriously by contemporary moral theorists, which can be the case only if the moral intuition that co-constitutes it – that is, (L) - is taken as sufficiently plausible. I will follow suit.

So how exactly does (L) help us to respond to the challenge mentioned above? There is an essential aspect of luck that could be described as a lack of control. This should be uncontroversial. Surely, an event that is under one`s control cannot be described as a *lucky* event. If that is so, then (L) could be reformulated in the following way:

> L*: An event that one has no control over cannot make a difference in one`s moral standing.

I take it that (L*) is no less plausible than (L).[223] Now, one`s mortality is clearly beyond one`s control and as such it belongs to the kind of events that cannot make a difference in one`s moral standing. In other words, death is not an *excuse* for wrongdoings that one would do if one didn`t die. Consider this:

[222] Thomas Nagel, `Moral Luck` in Daniel Statman (ed), *Moral Luck* (New York: State University of New York Press, 1993) pp.57-71, (p.58).

[223] However, I don`t wish to imply that *luck* and *lack of control* are synonymous concepts. They are not. There is going to be full moon in several days; an event that is totally beyond my control. To describe this as me being lucky (whenever there is full moon) would be rather weird. There is remarkably little discussion on the nature of the concept of luck among moral theorists. An intriguing exception is Nicholas Rescher, *Luck: The Brilliant Randomness of Everyday Life* (New York: Farrar, Straus and Giroux, 1995).

A terrorist plants a bomb in a theatre full of people. The bomb is controlled remotely. The terrorist contacts the authorities informing them that a bomb has been planted in an unknown public place. He presents a list of demands. After the negotiations with the authorities fail, the terrorist proceeds to detonate the bomb. He is about to push the button on the remote control when he suffers a sudden cardiac arrest. He drops the remote control before he manages to press the button; he passes out and a few minutes later he is dead.

It will be agreed, I believe, that the terrorist in our little story is an extremely wicked person. He will be seen as such despite the fact that he has caused no harm to anyone. He would have done it, had he not been prevented by his sudden death, and that is enough for us to judge him as morally corrupted.[224]

We can conclude that in the contexts of evaluating moral standings, one`s perishability is irrelevant. That is, in these contexts, a human agent must be seen as if she was imperishable. Above (section 5.4.2), it has been shown that (PP) won`t work for perishables. Thus, if man can be treated as imperishable in the contexts of evaluating their moral standing, then (PP) applies in those contexts.

5.4.8 For all men there is a time at which they do things that corrupt their moral standing

This follows from steps (1), (5) and (7) in the following straightforward way. (5) tells us that it is possible for any man to do things that corrupt their moral standing. (1) tells us that for some states of affairs, if they are possible then there is a time at which they obtain. And (7) tells us that doing things that corrupt one`s moral standing belongs to the state of affairs that (1) applies to. Step (7) is crucial here, as it allows to treat man as immortal. Without this step it could be objected that many men will simply die *before* they manage to do something that corrupts their moral stranding.

5.4.9 The moral standing of man is essentially corrupted

Above, we have concluded that for all men there is a time at which they do things that corrupt their moral standing. It seems obvious that a wrongdoing done in the past or in the present corrupts one`s *present* moral standing. It is much less obvious, however, that one`s *present*

[224] It could be pointed out that what makes us judge the terrorist as morally corrupted is his *intention* to do harm. This could be then taken as showing that it is not only future (or past or present) actions that corrupt one`s moral standing but the *intentions* to do them as well. This can be conceded without any harm to the logic of the argument. The reader is invited to understand an action as morally corrupting only if it is intentional.

moral standing gets corrupted by a *future* wrongdoing. Step (6) explains how that is the case nonetheless.

At this point we can draw something like the following picture of man`s moral standing. It is possible for any man to do a wrongdoing. There is a time at which all men will do a wrongdoing. Regardless of whether the time of one`s wrongdoing is in the past, in the present or in the future, the wrongdoing corrupts one`s (present) moral standing. Therefore, at any present moment, all men`s moral standing is corrupted. Another way of putting the last point is that: the moral standing of man is essentially corrupted.

5.5 Conclusion

We have come full circle. We want our practice of holding others responsible to be fair. The fairness of the practice depends on the possession of free will. The idea of possessing free will in a causally determined world constitutes a serious dilemma. Philosophers are concerned about the fairness of the practice and, therefore, they address the dilemma. A successful account of free will is expected to sustain the fairness of the practice. The philosophers addressing the dilemma work under the assumption that once the dilemma is solved there won`t be any other reason preventing us from pronouncing the practice as, in principle, fair. I submit, however, that there is another such reason. It is commonly accepted that the practice of holding another responsible is unfair if the holder`s moral standing has been compromised. I argue that morally compromised standing is essential to all men. Thus, the fairness of the practice cannot be saved even if a successful account of free will is formulated. This means that the fact that my compatibilist solution to the problem of free will doesn`t enable theoretical justification of the practice of holding others responsible shouldn`t be seen as a weakness of the solution. It shouldn`t be seen as such because the practice is – for a separate reason – essentially unjustifiable anyway.

Appendix

My argument here in Appendix will crucially depend on a philosophical interpretation of a passage from the New Testament. It is somewhat tricky to try to build a convincing philosophical argument around a passage from a religious text. Presumably, most of my readers will be either atheists and/or those unwilling to mix philosophical with theological assumptions. They will, therefore, see little theoretical value (regarding philosophical arguments) in appeals to religious texts. Thus, almost half of the chapter is devoted to convincing the presumably atheistic reader that a religious text can be relevant in a philosophical argument. The convincing will consist of two arguments. The first one relies on some rather controversial claims, and although I will provide an argument in support of each controversial claim, I do not expect the first argument to convince all or even most of my readers. I hope to have, in this respect, more success with the second argument as it employs considerably less controversial claims.

1.1 The epistemic status of a religious text: two arguments

I will extract the reasons for claiming that the moral standing of man is essentially corrupted from the following story in the New Testament:

> Then each of them went home, while Jesus went to the Mount of Olives. Early in the morning he came again to the temple. All the people came to him and he sat down and began to teach them. The scribes and the Pharisees brought a woman who had been caught in adultery; and making her stand before all of them, they said to him, "Teacher, this woman was caught in the very act of committing adultery. Now in the law Moses commanded us to stone such women. Now what do you say?" They said this to test him, so that they might have some charge to bring against him. Jesus bent down and wrote with his finger on the ground. When they kept on questioning him, he straightened up and said to them, "Let anyone among you who is without sin be the first to throw a stone at her." And once again he bent down and wrote on the ground. When they heard it, they went away, one by one, beginning with the elders; and Jesus was left alone with the woman standing before him. Jesus straightened up and said to her, "Woman, where are they? Has no one condemned you?" She said, "No one, sir." And Jesus said, "Neither do I condemn you. Go your way, and from now on do not sin again."[225]

I wish to extract philosophical reasons from a religious text. A secular, materialist reader will understandably look at such a wish with suspicion. Religious texts, almost by definition, draw their plausibility from an (alleged) divine authority. This is something that stands in a strong contrast to modern philosophy which – inspired by the Enlightenment – recognizes Reason and only Reason as the ultimate authority in theorizing about the world.

I am fully in agreement with this Enlightenment-inspired approach to philosophising. There is, however, something peculiar about moral issues that, at least in some cases, ties their discussion with relevant religious texts. There is an obvious, historical connection.

[225] *Bible* (*New Revised Standard Version*), John 7:53 – 8:11

Since the beginning of the Abrahamic faiths and of Greek philosophy, religion and morality have been very closely intertwined, in fact, inseparable until very recently. The truth of a moral claim used to be fully established by a reference to a relevant passage in a canonical religious text. In modern, predominantly atheistic, ethical theory, canonical religious texts don`t wield such authority anymore. They are, however, still seen, amongst other things, as records of what a culture is strongly inclined to take to be morally right and wrong. Thus, a modern (atheistic) moral theorist will respect the texts not as something that fixes the truth of their moral claims in virtue of being inspired by the divine but merely as something that has historically informed our moral axioms without providing, in itself, any resources to ground their truth.

Now, this obvious, historical connection between moral values and religious texts is not good enough for my little project of extracting philosophical reasons (relevant to a moral theory) from a religious text exactly because the texts themselves offer no resources to ground the truth of their moral claims. More is needed here. Below, I formulate and defend two arguments that purport to show that religious texts should be seen as carrying a significant epistemic weight in moral theorizing. The arguments won`t convince a moral non-realist as they assume the truth of moral realism. That shouldn`t bother us here because we have already lost a moral non-realist above (see sections 4.2 and 4.5.2.1) when we have tied the issue of justifying the practice of holding others accountable in a causally determined world to the notion of *fairness*. A moral non-realist has to resist such a tie because *fairness* is a notion that refers to an objective (mind-independent) *moral* property; that is, it refers to a *kind* of property that a moral non-realist doesn`t accept into their ontology.

1.1.1 The first argument

Before I give you an outline of the first argument, let me show my hand here as it will make it easier to see how the argument is supposed to work.

Atheism has always been associated with the worry that if God doesn`t exist then there are no moral obligations, no moral rights or wrongs. Richard Taylor, who is a non-theist, argues against moral realism on exactly these grounds:

> Our moral obligations can […] be understood as those that are imposed by God. […] But what if this higher-than-human lawgiver is no longer taken into account? Does the concept of moral obligation […] still make sense? […] [T]he concept of moral obligation [is] unintelligible apart from the idea of God. The words remain but their meaning is gone.[226]

William James seems to hold a similar view when he says:

[226] Richard Taylor, *Ethics, Faith, and Reason* (Englewood Cliffs, N.J.: Prentice-Hall, 1985), pp.83-84.

> The stable and systematic moral universe for which the ethical philosopher asks is fully possible only in a world where there is a divine thinker with all-enveloping demands.[227]

I wish to exploit the implication that God`s non-existence implies non-existence of moral obligations in my argument below. I will argue that being a moral realist commits one to the assumption of God. Under this assumption, the epistemic weight of religious texts increases dramatically. The following is, however, no knock-out argument. It rests on some highly contentious assumptions, and although I will try to dispel some of the potential objections, it will convince only some of my readers. The unconvinced will find more force in the second argument.

An outline of the argument:

E1: Moral realism: there are objective moral truths.

E2: The objective moral truths have a normative force.

E3: The force of a moral truth can be normative only if it comes from an external authority.

E4: That which is an (external) authority in relation to humans must itself be *trans*- or *super*-human, i.e. in some sense, higher in the ontological hierarchy.

E5: This trans-human authority might plausibly be seen as the God of religious texts.

E6: Recovering God as entailed in the theory of moral realism restores (for a moral realist) the epistemic authority of a religious text.

1.1.1.1 Moral realism: there are objective moral truths

(E1) is an assumption that has been implicit in the discussion already for some time (since section 4.2). Although the majority of philosophers are moral realists,[228] moral anti-realism is not a position that is obviously inferior to that of moral realism. There are good reasons to be a moral anti-realist.[229] (E1) explicitly stipulates the truth of moral realism, which immediately closes off the argument for an anti-realist. This, however, shouldn`t – as explained above – be seen as a weakness here. I lost a moral anti-realist already some time ago. In the context of grounding our practice of holding others accountable, free will is an

[227] William James, `The Moral philosopher and the Moral Life`, *International Journal of Ethics*, 1, no. 3 (1891), pp.330–354 (section V).

[228]PhilPapers, 'The Philpapers Surveys', ed. by PhilPapers (2014). <http://philpapers.org/surveys/results.pl?affil=Target+faculty&areas0=0&areas_max=1&grain=coarse> [accessed 21 March 2019].

[229] Perhaps the most plausible objection to moral realism draws its strength from the huge popularity of the evolutionary paradigm. The evolutionary paradigm provides powerful conceptual tools to interpret the so-called objective moral facts as, in fact, contingent adaptive mechanisms that vary with varying stages of our evolutionary development.

issue only if the practice is motivated by *desert/fairness*, and that is a condition that a moral anti-realist cannot accept.

1.1.1.2 The objective moral truths have a normative force

(E2): The objective moral truths have a normative force. Richard Joyce argues that normativity of moral truths is a *non-negotiable* commitment.[230] This should be uncontroversial. Morally right actions *ought to* be pursued, morally wrong actions *ought to* be avoided. Let me briefly deal with a possible, albeit a somewhat confused objection.

Hume famously argued that one cannot derive imperative statements from factual statements, that is, one cannot derive claims about what *ought to* be the case from what *is* the case.[231] Isn`t such a mistaken derivation what I am guilty of here? Don`t I first state there *are* moral facts, (E1), and then proceed to derive imperativness or normativity, (E2)? Well, not really. I agree with Hume: one can`t *derive* an `ought` from an `is`. I am not, however, *deriving* anything - I am *stipulating*. When the assumption of moral realism is introduced in (E1), *facts*[232] come inseparably mashed with *imperatives*. The peculiar nature of moral *facts* is that they *command*. Once something is *recognized* as a *moral* fact, it must be, at the same time, recognized as an *imperative*; that is, its normative force cannot go unrecognized. Eric D`Arcy takes this last point as self-evident in the strict sense: "If someone says, 'X is good,' it is nonsense to agree that it is, and to ask whether it is something that should be desired and pursued".[233] Thus, once the reader accepts that moral facts exist (as she should if worried about the *fairness* of our practice of holding others accountable in a world where alternative causal chains are impossible), she accepts the normative force that comes with them. That`s because (E1) entails (E2).

1.1.1.3 The force of a moral truth can be normative only if it comes from an external authority

(E3) is a crucial but, at the same time, rather controversial claim. The force of a moral truth is normative in the sense of it obliging an agent (who is confronted with it) to act in a certain way. Now, (E3) could be seen as combining two separate claims:

i. The normative force of a moral truth is, in some sense, out there in the world.
ii. Obligation entails authority.

[230] Richard Joyce, *The Myth of Morality* (Cambridge: Cambridge University Press, 2011), ch.1.

[231] The argument is located in the last paragraph of David Hume`s *A Treatise of Human Nature*, Book III, Part I, Section I.

[232] Although the claim in (E1) is formulated in terms of moral *truths*, it could be equally well (i.e. *salva veritate*) formulated in terms of moral *facts*.

[233] Eric D'Arcy, *Conscience and Its Right to Freedom* (New York: Sheed and Ward, 1961), p.53.

Most moral realists will be happy to endorse (i) while denying (ii). They will insist that the obliging force is *intrinsic* to a morally significant situation or, somehow, constituted[234] by the situation. An agent employs her sense of justice to detect the obliging force and acts accordingly. There is no need to posit an obliging *authority* somewhere in this picture. This position will be popular with those who are both moral realists and atheists. I appreciate the appeal and plausibility of the position: it is parsimonious, compatible with atheism and it seems to agree with the phenomenology of our moral experience. When involved in a morally significant situation, we *perceive* the obliging force as coming from *within* the situation, and not as something imposed on us by an authority. This is, I agree, a correct description of our moral experience. The question is whether the fact of such an experience *implies* that the *source* of the obliging force is to be found *in* that situation. It can be argued that it doesn`t. Consider the following simple analogy: A water tap above my kitchen sink is a *place* where my water comes from. At the same time, the tap is not the *source* of my water, which I am, at times and rather annoyingly, reminded of when the water supplier temporarily stops the supply for repairs or because of an emergency. Similarly, the morally significant situation might be where the obliging force is (spatiotemporally) *located* without the situation being its *source*.

The possibility of interpreting a morally significant situation as a mere (spatiotemporal) *location* and not as a *source* of its obliging force represents an alternative that some may find attractive. Who and for what reasons? In principle, a moral non-realist should accept the (theoretical) distinction between a morally significant situation as a *location* and as a *source* of its obliging force as such a distinction is compatible with her (typical) endorsement of the is-ought gap. A moral non-realist can then appeal to the idea that as a mere *location*, a situation doesn`t intrinsically involve any obliging force. It would do so only as a *source* of it. And, of course, a moral non-realist will deny that situations can be such *sources*.

What about a moral realist? There will be at least two groups of moral realists who might be happy to accept that a morally significant situation is a mere *location* and not a *source* of the obliging force.

The first one is a group of Kantians. For a Kantian, the obliging force has its *source* in the autonomy of a rational agent and not in a morally significant situation. Thus a Kantian will find the distinction between a situation as a *location* and as a *source* of the obliging force

[234] A common strategy is to appeal to the theoretical notion of supervenience here. The notion of supervenience applied in ethics gives us something like the following necessary connection between natural facts and their moral properties: there cannot be any moral difference between two possible states of affairs or actions without there being some difference regarding the natural facts that constitute the respective states of affairs or actions. This claim is widely accepted and as such rarely argued for. An exception (regarding an attempt to argue for it) is Michael Smith `Does the Evaluative Supervene on the Natural?`, in Michael Smith, *Ethics and the A Priori*, (Cambridge: Cambridge University Press, 2004), pp.208–233.

useful as it allows her to explain certain phenomenological aspects of our moral experience – for instance, the impression that the obliging force is somehow *in* the situation - while retaining her claim that the actual *source* of it is in the autonomy of a rational agent.

The second group consists of moral realists who – while insisting that the *source* of the obliging force lies in the morally significant situation – wish to have a `plan B` in reserve. Why should a moral realist of the second group wish to have a `plan B`? The answer has to do (again) with the distinction known as *is-ought gap*. The distinction is commonly formulated in terms of an implication: the fact that something *is* the case doesn`t *imply* that something *ought* to be the case. This could be paraphrased in the following way: the empirical facts constituting a situation never *imply* an obligation to act in a certain way. A moral non-realist sees this as posing a serious problem for a moral realist: if what *is* cannot imply a (moral) *ought* then in what sense can the *oughts* be understood as existing at all? And if there are no *oughts* out there then, perhaps, there are no objective moral facts out there either.

A moral realist can dismiss this little argument here as inconsequential regarding her position in, roughly, the following way: True, there is no relation of *implication* between the *factual* statements and the *imperative* statements; that is, one cannot *derive* an `ought` from an `is`. However, a moral realist rejects the view that an `ought` is something that is essentially *derived* from something else. A moral realist sees `oughts` as fully given in the factuality of a situation. As such, the `oughts` are not *derived* but *perceived* or *recognized*.

Now, it should be noticed that for a moral realist a lot is at stake here. If, somehow, her account of how exactly the `oughts` are fully given in the factuality of a situation is unconvincing, then the plausibility of moral realism as a whole will suffer. At this point, a moral realist might welcome the `plan B`, i.e. the theoretical benefits of the distinction between a situation as a *location* and as a *source* of the obliging force. The distinction will allow her to explain how an obliging force can be associated with a situation without the situation being a *source* of it. This move might help her to save the plausibility of moral realism. (The move comes at a price though. The moral realist either comes over to the Kantian camp[235] or follows my argument at the end of which the source of the obliging force gets located in God.)

The distinction between a situation being a mere (spatiotemporal) *location* and being a *source* of the obliging force is a crucial one. As argued above, it is a distinction that might be acceptable only for some philosophers: most moral non-realists won`t protest, I expect, and the two groups of moral realists too should find the distinction useful. The reader will remember that we have lost a moral non-realist already some time ago and therefore I shall,

[235] Which I shall try to dissuade her from below.

in the rest of the argument, address the two groups of moral realists only. Others, that is those who reject the distinction, can skip the rest of the discussion of the first argument and proceed to section 1.1.2 where the exposition of the second argument begins.

1.1.1.3.1 A Kantian response

Let`s assume then that it has been conceded that the source of the obliging force cannot be *in* the situation. Let`s also assume that (ii) is correct: obliging entails authority. Do those two assumptions give us (E3)? Not necessarily. It is possible to take a Kantian way and argue that the source of the obliging force is in the self-authoritative structure of agent`s autonomy. Kant argues that we all are autonomous rational beings capable of willing and doing the rational. This particular combination of being an autonomous plus rational plus agential being is what constitutes an obligation we have towards ourselves (and to others to the extent to that they are autonomous agents) to act rationally. (The details of how exactly Kant conceived of the obligation being constituted by that combination are controversial.[236] Here I will assume, for the sake of the argument, that possessing autonomy can be plausibly understood as a source of obligation to act in certain ways.) In other words, being autonomous rational agents obliges us to conform our actions to rational principles of conduct that we (as rational beings) accept independently of desire. Kant then proceeds to formulate the ultimate rational principle that all autonomous rational agents are obliged to conform to in their actions. He calls this ultimate rational principle *The Categorical Imperative*. It reads as follows: `Act only according to that maxim whereby you can at the same time will that it should become a universal law`.[237]

Kant seems to assume that there is a connection between being obliged (as an autonomous being) to act rationally and the Categorical Imperative, such that we are similarly obliged to act as the Imperative dictates. I am unable to see how we can get from the law of an autonomous rational will (i.e. from the Kantian concept of autonomy) to the specifics of the Categorical Imperative. Here I am in total agreement with Thomas Hill who takes this transition `from an undeniable formal principle to a dubious substantive principle`[238] to be illegitimate. Neither is this, however, the point I want to press. I have no reason to rule out that the transition is, in principle, possible and thus I prefer to challenge the Kantian response to (E3) at a different point.

[236] Perhaps the most promising attempt to reconstruct Kant`s argument in a way that would make things here clear and immune to some of the objections raised against it is Christine M. Korsgaard, *The Sources of Normativity* (Cambridge: CUP, 1996), ch.2,4.

[237] Immanuel Kant, *Grounding for the Metaphysics of Morals*, James W. Ellington (trans.) (Hacket, 1993 [1785]), p.30.

[238] Thomas E. Hill, Jr., *Dignity and Practical Reason in Kant`s Moral Theory* (Ithaca: Cornell University Press, 1992), p.122.

Once we ignore the afore-mentioned problems, we are able to lay the Kantian response down in roughly this way: As autonomous rational agents, we are obliged to act rationally. The ultimate principle of (practical) rationality is the Categorical Imperative. Thus, we are obliged to act as the Categorical Imperative dictates. An application of the Categorical Imperative identifies some actions as those an autonomous agent is obliged to pursue, others as those to be avoided. In this way, the obligation – that is, *normativity* – has its source in the autonomy of the agent and not, in some sense, outside of it. This is a moral theory, thus the domain that the Categorical Imperative helps us to navigate is the domain of moral truths. The normativity of those moral truths traces back to the autonomy of the agent facing them. So far so good. However, cracks start appearing once we turn our attention to an assumption behind this picture of moral normativity. The assumption here is that *the structures of the moral and the rational domains are mutually isomorphic*. That is, it is always morally right to do what the Categorical Imperative (as the ultimate principle of practical rationality) dictates, and it is never morally right what the Categorical Imperative doesn`t dictate.

A Kantian about moral normativity must be committed to this assumption because should it be the case that either something is morally right despite not being identified as such by the Categorical Imperative or something is identified by the Categorical Imperative as morally right despite it clearly not being such, then it can`t be maintained that normativity associated with moral truths has its origins in agent`s autonomy. Once the alleged normative link starting in the autonomous agent and running via the Categorical Imperative towards the moral domain can be (in some cases) severed, it will have to be concluded that the moral truths possess their normative force independent of the agent`s autonomy.[239] And severed it can be.

First, let`s have a look at how the Categorical Imperative is meant to work. Suppose you wonder whether you ought or oughtn`t steal from another (to become an owner of something you desire). Now run the two possible answers through the decision procedure of the Categorical Imperative. The Categorical Imperative – being the ultimate principle of (practical) rationality – dictates to look for contradictions when a maxim is applied. Thus, we shall ask: does any of the two lead to a contradiction if universalized into a moral law as dictated by the Categorical Imperative? It looks like one of them does. Test the positive answer and imagine a world in which `you ought to steal from another` is made into a universal moral law. In such a world the concept of *private* property becomes incoherent. If

[239] There is a sense in which the normativity of the moral depends on autonomy. If normativity is understood as a force that obliges one to do something then, of course, it can do so only if one can be made obliged. A non-autonomous being – i.e. a being incapable of willing freely – cannot be made obliged. The moral has no normative force relative to a non-autonomous being. This, however, doesn`t translate into moral normativity having its source in the autonomous agent. Autonomy is a necessary condition for moral normativity, not its source.

stealing the property of another is *right*, then in what sense is that property private? Arguably, something is someone`s private property only if it is *rightfully* hers. But something cannot be *rightfully* someone`s in a world where it is, at the same time, *right* to take that property from the *rightful* owner. That`s a clear contradiction. Testing the negative answer won`t lead to a contradiction (I will leave it to the reader`s imagination to test this claim) which makes it the *right* answer.

A similar case – discussed by Kant[240] – would be wondering whether it is morally right or wrong to make a promise while having no intention of keeping it (in order to get needed money). Now, again, it looks like we run into a contradiction if we approve of such an action and subject it to the logic of the Categorical Imperative. In a world in which the right thing to do is to make a promise with no intention of honouring it, the very concept of *promise* becomes incoherent. The concept of promise is closely related to the concept of trust; they are part of the same transaction. The act of promising (something to someone) is successfully accomplished only when complemented by a related act of trusting (the promise). But in such a world, the logic of the maxim – the maxim that dictates that making false promises is the right thing to do – turns trusting into an irrational act. One cannot trust in such a world. And, consequently, one cannot successfully accomplish the act of promising.

So far so clear. However, scenarios can be conceived of in which the application of the Categorical Imperative won`t, at least on the face of it, yield intuitively correct results. I will briefly consider two such scenarios. In one of them you are a doctor with a newly born child in your care. The child suffers from a rare disease that will require near constant medical attention for the rest of her life. You ask yourself: `Should I euthanase this child (now when she is still unconscious) so the NHS saves money for her treatment?`. In the other scenario you wonder: `Should I become a politician to help steer the country off its disastrous political path?`.

Regarding the first scenario, it would, clearly, be a morally wrong thing to euthanize the newborn. At the same time, I am quite unable to universalize it in some such way that it would lead to a contradiction. We can easily conceive of a society that approves of euthanizing incurably ill newborns on the grounds that their treatment would be too costly. Each year there would be certain percentage of newborns euthanized under those circumstances and the money thus saved would be spent on life saving operations for kids with good prospects of full recovery.

Although, as admitted above, I can`t think of a contradiction resulting from a universalisation of the maxim that governs the first scenario, I don`t want to rush to

[240] See Immanuel Kant, *Grounding for the Metaphysics of Morals*, James W. Ellington (trans.) (Hacket, 1993 [1785]), p.31.

conclusions here. Others might be able to uncover a contradiction here that I don`t see and save the reliability of the Categorical Imperative.[241] There seem to be different ways of interpreting the concept of `contradiction` (involved in the application of the Categorical Imperative), which enable very creative ways of uncovering contradictions in universalisations of maxims that would, on the face of it, seem as hopeless targets of such uncovering.[242] What I, instead, wish to point to here is this: Many similar scenarios that present this kind of challenge can be conceived of. A creative thinker will perhaps be able to find a way of recovering a contradiction in the universalisation of the relevant maxims in all of these scenarios. But the more scenarios like this we present as a challenge to the creative Kantian, and the more she successfully interprets them as leading to a contradiction, the more difficult it will become not to notice that in her quest to save the Categorical Imperative, it is mostly the logic of the Categorical Imperative and not our moral truths that gets adjusted. Always much more – if not all – of the adjusting activity will target the logic of the Categorical Imperative than our moral truths. This strongly suggests that our acquaintance with and commitment to certain basic moral truths comes before our acquaintance and commitment to anything like the Categorical Imperative. This (ontological) priority and independence of moral truths with respect to the Categorical Imperative doesn`t sit well with the claim that the normative force of moral truths traces back - via the Categorical Imperative – to the agent`s autonomy. More about it later. Let me turn to the other scenario now.

Concerned about the rise of far-right sentiments in your country, you might wonder: ` Should I become a politician to help steer the country off its disastrous political path?`. That`s a perfectly understandable, self-directed question. Becoming a politician for the above-mentioned reason is commendable. However, it becomes immediately obvious that when the relevant maxim is universalised it will result in a (practical) contradiction. A country in which everyone becomes a politician to battle the problems in the society will nearly immediately collapse as there won`t be any workforce available to sustain the functioning of non-political institutions. The problem then is this: intuitively, we would approve of someone

[241] Some could appeal to Kant`s second formulation of the Categorical Imperative that commands that one should always treat one`s own or other person`s humanity as not simply a means but always at the same time as an end. Such an appeal should, on the face of it, give a correct result, and save the newborn. There are two problems here though. First, the second formulation of the Imperative is much further removed from the plausible formalism of the law of autonomy and the first formulation of the Imperative. It is simply too morally laden and as such it itself requires a justification, which cannot, I believe, be given within the framework of Kant`s moral theory. Second, the notion of humanity appealed to in the second formulation is essentially connected to that of autonomy. It is unclear, however, to what extent a newborn possesses any autonomy and, consequently, any humanity.

[242] For an illuminating discussion of the different kinds of contradictions compatible with Kant`s discussion of the Categorical Imperative, see Christine Korsgaard, `Kant`s Formula of Universal Law`, *Pacific Philosophical Quaterly,* 66, no. 1-2 (1985), pp.24-47.

wanting to become a politician for altruistic reasons, yet the Categorical Imperative informs us that such an action would be immoral.

Now, an obvious and perhaps even correct reply would be to point out that choosing a career path is not (even if motivated by altruistic considerations) a *moral* action. And if it is not a *moral* action then there is no *moral value* that the Categorical imperative can help us to determine. Let`s concede, for the sake of argument, that the reply is correct and the Categorical Imperative is not applicable here. The problem with this reply in the context of claiming that the normative force runs from the autonomous, rationally willing agent - via the application of the Categorical Imperative - towards the moral domain is this: Imagine you wish to determine whether a particular action would be morally right or wrong. You intend to apply the logic of the Categorical Imperative to accomplish that. You are aware of the constraints of application of the Imperative – it gives correct results only if applied in *morally* significant contexts. At this particular point, the question is: how do you know which contexts are *morally* significant and which aren`t? The Categorical Imperative won`t tell you that. Its correct application already presupposes that you are able to distinguish the contexts. This, if true, strongly suggests that our access to the realm of moral truths comes prior to and independently of the Categorical Imperative.

Having taken a different path, we arrive at a very similar conclusion here to the one arrived at above when we discussed the scenario of a doctor wondering whether to euthanize an incurably ill infant: moral truths seem to be out there prior to and independent of the Categorical Imperative. If this is so then the normative force of the moral doesn`t trace back to an autonomous agent. If the normative force can be detected out there in the moral domain in those cases in which the connection between the autonomous agent and the moral domain cannot be established via the application of the Categorical Imperative, we have to conclude – for reasons of parsimony – that the normative force is coming from outside the agency.

Let me conclude the discussion of (E3) by coming back to the quote by Richard Taylor. He says, among others, that `[T]he concept of moral obligation [is] unintelligible apart from the idea of God. The words remain but their meaning is gone` (see p. 130). At this point, I am not able to confirm that the intelligibility depends on the assumption of God; that will be argued for later. I believe that the intelligibility depends on the assumption of `higher-then-human lawgiver`, and I see this as a default position. It entails the truth of (E3) because the concept of `higher-then-human lawgiver` refers to an authority. Kant`s moral theory – particularly his concept of autonomy – seems, on the face of it, to provide theoretical resources to challenge the default position as it promises to derive all normative force from the autonomously willing agent. If the bounding or normative force of duty (to do the morally right thing) can be construed as derived from the normative force of the Kantian autonomy, then we should (as parsimony dictates) do without the assumption of a `higher-than-human lawgiver` (or God). However, closer scrutiny reveals that moral truths are – as far as their normativity is concerned - essentially independent of the agent`s autonomy and as such cannot be taken as deriving their normative force from it. This brings us back to (E3), that is, to the default position.

1.1.1.4 That which is an (external) authority in relation to humans must itself be trans- or super-human, i.e. in some sense higher in the ontological hierarchy

I suspect that (E4) is one of those claims that are difficult to argue for in virtue of it expressing a rather simple conceptual truth that is not easily broken into simpler, constitutive claims. Uncontroversially, `authority` refers to, among others, the power or right to oblige others to act in a certain way. Having this kind of power or right entails a hierarchy of some sort and the authority having a higher position in that hierarchy in some relevant sense. A normal mature human being can, in most circumstances, detect (via her sense of justice) the obliging force of a moral fact that she is confronted with. An (external) authority reaches to us through this obliging force. Now, once we posit humans as a *whole* as the subject of an obliging authority, we must conclude that such an authority must be *trans*- or *super*-human in some relevant, plausibly ontological, sense.

1.1.1.5 This trans-human authority might plausibly be seen as the God of religious texts

A being that is (a) ontologically superior to humans and (b) possesses the power to morally oblige all of the humankind is, plausibly, the God of religious texts.[243] Now, while there might be other ontologically superior beings out there without them being God(s), it is somewhat difficult to see how our concept of God could survive placing the source of the morally obliging force into some other being. That would amount to saying that God is not the ultimate authority, because if He is not the ultimate authority regarding the moral then in what sense is he an *authority* at all? And an *authority* He must be if He is the God.

There is a well-known objection to the claim that God is the normative source of the moral. The objection has its origins in what has come to be known as the *Euthyphro dilemma*. In *Euthyphro*, one of Plato`s dialogues, Socrates asks Euthyphro: `Is the pious loved by the gods because it is pious, or is it pious because it is loved by the gods?`[244] Now, although the original dilemma is about a relation between the pious and gods, it can easily be modified to be about a normative relation between the moral and God. One such modification was done by Leibniz:

[243] I will presume here that religious texts of various religions refer, ultimately, to the same God. (This, of course, presupposes Millianism about reference, which is fine with me.)
[244] Plato, *Euthyphro*, Benjamin Jowett (trans.) (Adelaide: Adelaide University Press, 2014), 10A

> It is generally agreed that whatever God wills is good and just. But there remains the question whether it is good and just because God wills it or whether God wills it because it is good and just; in other words, whether justice and Goodness are arbitrary or whether they belong to the necessary and eternal truths about the nature of things.[245]

The dilemma poses two problems for my argument. First, it shows that it is far from inconceivable to see the moral as being normatively independent from God (God being, in fact, in some sense a subject to the external normative force of the moral). Clearly, agreeing to this normative independence amounts to giving up (E4). I could, perhaps, insist that such a normative independence is not really conceivable and that the moral has its normative force in virtue of being willed by God. This reply, however, leads to the second problem: if the moral value of facts derives from God`s will, it follows that moral truths are *arbitrary* (as God could, surely, have willed wildly differing sets of moral truths). The reader will remember that the whole first argument has been predicated on the assumption of moral realism. Thus, the arbitrariness implied in my answer to the first problem clashes with the assumption of moral realism, which, of course, would collapse the whole argument.

Now, the reader won`t be surprised to hear that each horn of the dilemma can be associated with a long list of advocates. Generally, philosophers who are rationalists and/or Platonists and/or realists about the moral will tend to defend the view that God issues commands *because* they are right.[246] The divine command theorist or voluntarist, on the other hand, will insist that what makes something right is the act of God commanding it.[247] As mentioned above, either horn of the dilemma seems to be incompatible with my argument: If God commands something *because* that something is right, I won`t be able to place the source of moral normativity in God. If, on the other hand, something is right *because* God commands it then it looks like I will have to give up moral realism (due to moral truths turning out to be arbitrary), that is, I will have to give up a key assumption of my argument.

However, the Euthyphro dilemma is, thank God, a false one. It projects a dubious Platonist intuition about the realm of *eidos* onto the concept of God. According to Plato, that which is *real* is located in the world of its own, a world that is ontologically independent of gods. Within this Platonist framework, one is a moral *realist* only if she places the moral

[245] Gottfried Leibniz, *"Reflections on the Common Concept of Justice"*, in Leroy Loemker, *Leibniz: Philosophical Papers and Letters* (Dordrecht: Kluwer, 1989 [1702(?)]), pp.561–573 (p.516).

[246] Philosophers endorsing the position are, for instance, Socrates, Averroes, Gabriel Vasques, Grotius, Leibniz, Ralph Cudworth, Samuel Clarke, Richard Price. Contemporary philosophers embracing this horn of the dilemma are Richard Swinburne – see his *The Coherence of Theism* (Oxford, Clarendon Press, 1993), pp.209-215 – and Tim J. Mawson – see his "The Euthyphro Dilemma", *Think*, 7, no. 20 (2008), pp.25–33.

[247] The list of philosophers supporting this position is similarly impressive to the one associated with the first horn of the dilemma. Some of the names on the list include Duns Scotus, William of Ockham, Martin Luther, John Calvin, Descartes, Thomas Hobbes and, perhaps somewhat surprisingly, Wittgenstein (I shall return to Wittgenstein below and be slightly more specific about his view).

truths into this realm of *eidos*. Placing them there is, at the same time, disconnecting them – in certain relevant sense discussed below – from God(s). However, this Platonist separation (and subordination) of God(s) from the world of *eidos* not only feels like a rather arbitrary constraint on God`s absoluteness in all aspects, it stands in clear contradiction to classical Judaeo-Christian theism. The classical Judeo-Christian theism would find such a (schizophrenic) split within the transcendental not only as seriously heretic but as deeply puzzling too. The Church Fathers, and others after them,[248] took the moral to be part of God`s essence. Thus, God doesn`t subordinate Himself to anything external when willing the moral, neither is He being *arbitrary* about the content of His moral commands. The commands are externalizations of His *essence*, and as such they are *necessary*. Katherin A. Rogers writes:

> Anselm, like Augustine before him and Aquinas later, rejects both horns of the Euthyphro dilemma. God neither conforms to nor invents the moral order. Rather His very nature is the standard for value.[249]

Once it is accepted – and accepted it should be, I believe – that the Euthyphro dilemma is a false one (because it relies on a very dubious and counterintuitive conception of God), I can go on claiming that the moral gets its normative force from God, who commands it. The claim doesn`t come at the cost of losing the moral realist. In the act of commanding the moral, God externalizes His own essence. God`s essence is *real* and *necessary*, and, in conceptual terms, it provides all that a moral realist requires.

I am fully aware of being very sketchy about a very deep issue here. I would have to deviate considerably from the main line of my argument to hope to make (E5) more convincing against the challenge of the Euthyphro dilemma. And even a lengthy deviation wouldn`t guarantee a success. The nature of the issue calls for a *theological* discussion, which wouldn`t be perceived as particularly convincing by most contemporary philosophers. I will instead, in a last-ditch effort, appeal to an authority here. In a letter from December 1930 to Friedrich Waismann, Wittgenstein said the following:

> Schlick says that in theological ethics there are two interpretations of the Essence of the Good. On the shallow interpretation, the Good is good, in virtue of the fact that God wills it; on the deeper interpretation, God wills the good, because it is good. On my view, the first interpretation is the deeper: that is good which God commands. For this blocks off the road to any kind of explanation, `why` it is good; while the second interpretation is the shallow, rationalistic one, in that it behaves 'as though' that which is good could be given some further foundation.[250]

[248] Among them Thomas Aquinas, William James, Wittgenstein.

[249] Katherin A. Rogers, *Anselm on Freedom* (Oxford: OUP, 2008), p.8.

[250] Quoted in Allan Janik and Stephen Toulmin, *Wittgenstein`s Vienna* (New York: Simon&Schuster, 1973), p.194.

To those unconvinced by my sketchy dismissal of the Euthyphro dilemma and Wittgenstein`s view on the issue, I promise a different argument towards the epistemic significance of religious texts below (after the discussion of step E6). Let me turn to (E6) now.

1.1.1.6 Recovering God as entailed in the theory of moral realism restores (for a moral realist) the epistemic authority of a religious text

Let`s assume that moral realism implies the existence of God. How does this assumption affect the epistemic significance (regarding moral issues) of relevant religious texts?

Consider, first, what the epistemic status of religious texts would be in a world without God. Arguably, with God out of the picture, the epistemic status of religious texts would derive from the epistemic status of its authors, that is, from the epistemic status of other human beings. I am unable to see how that would work. First, we don`t have any evidence that the authors of religious texts possessed superhuman minds (or souls) that were so vastly superior to the rest of humanity in their (and future) time that the (moral) claims they made should be attributed a very special status: a status that is obliging the rest of humanity in virtue of it having been authored by a kind of superhumans. And even if we had some such evidence about the super-human status of the authors, we could either dismiss it as not superior enough (because not *absolute*) to sustain moral realism, or we could take it as sufficient to sustain moral realism and conclude that the epistemic status of religious texts (in matters regarding the moral) is satisfactorily high.[251]

Second, the texts themselves rarely, if at all, contain any argument for their claims. If the epistemic status of the texts cannot be derived from the epistemic status of its authors and the claims are unargued for, then the status must be seen as extremely low, if any. Thus, it appears that in a world without God the religious texts can hardly be taken as an authoritative source of knowledge about the moral.

Now, what if God does exist? Would it be a good reason to take religious texts seriously in matters of the moral? I think so. Our concept of God is given its content by, among other things, the relevant religious texts. Not taking the texts seriously amounts to accepting that we don`t really have a concept of God. At best, we only have an empty conceptual shell. Of course, this is possible. There is no way of silencing the sceptic here. We might be fundamentally wrong and confused about anything we believe to know about God. The relationship between God and man is tremendously complicated in virtue of man having been created as *free*, i.e. as able to deviate from paths approved of by God. The wicked ones might

[251] Although I don`t see how a lesser than absolute authority could oblige all of the mankind, someone might be able to give a plausible account. In that case, I would ground the epistemic status in the *super* status of that other, lesser than absolute, authority. The logic of my argument would stay unaffected.

have burnt the true, god-inspired texts replacing them with texts containing only lies. Thus, perhaps the only thing we know is that He is "that than which nothing greater can be conceived".[252] But this cannot be taken as a default position. It is the position of a sceptic, which is a position that cannot be a default position because it is not a position at all: it is an absence of a position. The default position in the world in which we take God to exist is the one in which He revealed Himself in the texts of god-inspired authors: the religious texts. Casting doubts on that position is easy but it doesn`t have much force if not complemented by a plausible alternative. And I don`t think there is one.

I don`t expect many will find the argument convincing. A Kantian will refuse to give up autonomy as the source of moral normativity, a Platonist will insist that true moral propositions inhabit a realm that is ontologically independent of God(s), and a sceptic will point out that the alleged connection between God and religious texts is just a myth. To them, I offer a different argument.

1.1.2 The second argument

Let me give you an outline of the argument first. The outline will be followed by a discussion of its individual steps.

F1: The Rawlsian assumption: we have a sense of justice.

F2: The sense of justice allows us to distinguish *right* from *wrong*.

F3: Within the Christian paradigm, the New Testament is the foundational text regarding the moral realm. It has been long *recognized* as a text that is *correct* about moral facts.

If we – beings with a sense of justice capable of distinguishing right from wrong – recognize the New Testament as a text that is correct about moral facts, then we should conclude:

F4: The New Testament *is* correct about moral facts.

1.1.2.1 The Rawlsian assumption: we have a sense of justice

In section 4.6.2.1, I have briefly discussed Rawls` concept of a sense of justice. Let`s assume there is no internal incoherence in the concept itself. Still, what reasons do we have to accept that we do, as a matter of fact, possess some such capacity that the concept of a sense of justice refers to? To answer the question, I will briefly sum up Rawls` argument here, and offer a supporting consideration of my own.

[252] Saint Anselm, *Proslogion*, M.J. Charlesworth (trans.) (London: University of Notre Dam Press: 1979), ch.2.

Rawls starts off his essay `The Sense of Justice` with Rousseau`s claim `that the sense of justice [is] the natural outcome of our primitive affections`.[253] Rawls proceeds to `set out a psychological construction to illustrate the way in which Rousseau`s thesis might be true`.[254] The construction draws heavily on Jean Piaget`s work, *The Moral Judgement of the Child*.[255] At the heart of the construction, as I understand it, is the following psychological observation. In normal circumstances, parents love their child. In time, the child develops `an equal love for the parent[s]`.[256] The relation of love towards one`s parents gives rise to a capacity to feel guilt after violating the precepts or injunctions issued by the parents. The child feels guilty because she recognizes and experiences the violation `as a breach of the relation of love and trust with the authoritative person`.[257] The feeling of guilt towards the loved parents is painful and the child seeks reconciliation and restoration of the previous relation. The child is able to achieve this reconciliation and restoration only if she is able to apply to herself the external (that is, in this case, coming from the parents) standards of criticism. This ability to apply to herself the external standards of criticism constitutes the grounds for, later in life, being able to form co-operative relationships of love and trust with others; `the participants [in a co-operative partnership] are bound by ties of friendship and mutual trust, and rely on one another to do their part`.[258] All this is possible only because the participants have a deep and formative experience of loving and trusting the other (initially, the parents) and an ability to conform to external (and universal) standards of cooperation. Such cooperation sustains the friendship and trust among the participants only if everyone `fulfil[s] their duty of fair play`.[259] The duty of *fair* play requires each participant to develop a `sense` for *fairness*, that is: *a sense of justice*.

The crucial point to notice here is that a sense of justice is a capacity whose developmental emergence is fundamentally grounded in our hardwired need to form a loving and trusting relationship with our parents. To form such a relationship with our parents is, as a matter of fact, how we are as species psychologically constituted. Of course, the psychological reconstruction (offered by Piaget and Rawls) of how this psychological constitution of ours develops into the sense of justice can be questioned. Still, it does seem to agree with the common knowledge of how children develop the social skills of interacting and cooperating with others, and as such represents a plausible account of how we come to possess something like a sense of justice.

[253] Rawls, `The Sense of Justice`, p.281.
[254] Ibid.
[255] Jean Piaget, *The Moral Judgement of the Child* (London: Kegan Paul, Trench, Trubner & Co. Ltd., 1932).
[256] Rawls, `The Sense of Justice`, p.287.
[257] Ibid.
[258] Ibid., p.289.
[259] Ibid.

To those who find the reconstruction implausible, I offer the following alternative consideration. Suppose you are a moral realist who, for some reason, doesn`t believe that we possess anything like a sense of justice. As such, you will face a rather problematic dilemma. There are moral facts or truths out there, but we either don`t have *any* access to them, or we have only a *rational* access to them. The former would be a rather unattractive implication. Somehow, we know there *are* moral facts, but we don`t know *what* they are. I am unable to see why anyone would want to be a moral realist and a moral agnostic at the same time. One of the points, I take it, of being a moral realist is to be able to search for and formulate true moral propositions. Being an agnostic moral realist is a bit like being a horse-riding instructor on a planet with no horses. What about taking the Kantian path claiming that we do have access to moral truths but this access is not of a *perceptual* but of a *rational* kind? The answer to this question depends on whether one believes (or not) that the moral and the rational are structurally isomorphic. That is, it depends on whether one believes that a competent application of the Categorical Imperative (or some such universal rational principle) always yields a morally *right* answer. Above (section 1.1.1.3.1), I argue that the Categorical Imperative is a rather unreliable pathfinder in the realm of the moral. Our ability to spot the unreliability should be understood as an evidence of us having a different – presumably a *perceptual*[260] – access to the moral realm. Thus, the Kantian position is – despite being an attractive one - unsustainable.

1.1.2.2 The sense of justice allows us to distinguish right from wrong

Above (section 4.6.2.1, p. 105), a `sense of justice` is described as `a capacity of persons to detect whether a state of affairs is or isn`t just/fair`. Now, although I believe that what is being predicated about the sense of justice in (F2) is interchangeable with how it gets defined above on p. 7, this is not something that is immediately obvious. The issue here is whether being able to detect whether a state of affairs is just/fair is the same as being able to distinguish right from wrong. (Or, at least, the issue is whether the ability to detect fairness/justice entails the ability to distinguish right from wrong.)

[260] The reader might remember that in section 4.6.2.1, I talk about moral truths as truths that are not `empirical` in nature and as such they are `not given in perception`. In this section, however, I describe our access to moral truths explicitly as `perceptual`. This apparent contradiction can be explained away in the following way. In section 4.6.2.1, my point was that moral truths or facts are not sense data in the same way as the sense data that we acquire via our five basic senses: touch, sight, hearing, smell and taste. Moral truths are not something that we have access to via those five basic senses; and it is in this sense that moral truths are not given in perception. The concept of perception can, however, be applied in a broader context, i.e. in a context that goes beyond the five basic senses. This broader context is a context which takes (various kinds of) *intuition* to be a *perceptual* capacity of ours even though this capacity cannot be – in any obvious way – associated with any of the five senses. The notion of *the sense of justice* refers to a capacity that is perceptual in this broader sense, that is, to a capacity that is essentially *intuitive*.

There are some transparent cases where the interchangeability is uncontroversial. `To execute an innocent person is *wrong*` seems to mean the same as `to execute an innocent person is *unfair*`. But what about the following pair of moral statements: `to refuse to jump into a pond to save a drowning child is *wrong*` and ` to refuse to jump into a pond to save a drowning child is *unfair*`? Surely, the latter doesn`t sound right. Agreed, it doesn`t sound right.

I offer the following as a way out. Above (p. 105), I define *fairness* in terms of *moral equilibrium*. More specifically, I define something as (being) *fair* as that something (being) *in moral equilibrium*.[261] I take it as uncontroversial that: it is *wrong* when the moral equilibrium is disrupted; i.e. it is *wrong* when something is *unfair*. The question now, of course, is whether the situation in which one refuses to jump into a pond to save a drowning child constitutes a situation in which the moral equilibrium is disrupted. If it does, then it is *wrong* because it is *unfair*; and vice versa, because things become unfair only due to a *wrongness* taking place. So, what is the answer to the question?

The notion of a moral equilibrium is suitably abstract and accommodating. As such, it can be used to express a moral evaluation of, I believe, any morally significant situation. In the case of a person who refuses to save a drowning child we could, for instance, say something like the following: Things are in moral equilibrium if one helps (within certain constraints regarding the availability of means and the potential danger to oneself associated with helping) a person in distress. Refusing to save a drowning child is refusing to help a person in distress. Consequently, the moral equilibrium gets disrupted.

There is, however, an issue here I have to briefly deal with. I treat the proposition that `things are in moral equilibrium if one helps a person in distress` as implying that `one *ought to*, morally, help a person in distress`.[262] There is a familiar objection to this kind of implication. Some acts are claimed to be such that they would be good to do, but not wrong not to do. Such acts are called *supererogatory*.[263] Consider the following:

The core of a nuclear reactor is melting. A team of emergency technicians has to be deployed to perform an emergency shut down to prevent a radioactive contamination of a densely populated area. Unavoidably, the team will be exposed to lethal levels of radioactivity.

[261] See footnote 176.

[262] From `things are in moral equilibrium if one helps a person in distress` I conclude that not helping a person in distress results in `the moral equilibrium get[ting] disrupted`. Disrupting the moral equilibrium is *wrong*, and we are obliged (ought to) avoid doing *wrong* things. Thus, one *ought to*, morally, help a person in distress.

[263] The concept of a supererogatory act was first introduced and discussed by James O. Urmson in his "Saints and Heroes", in Abraham I. Melden (ed.), *Essays in Moral Philosophy* (Seattle: University of Washington Press, 1958).

Clearly, it is a (extremely) good thing to shut down the reactor, yet it doesn`t feel correct to say that the team of technicians is morally *obliged* to sacrifice their lives in the act of saving lives of others. Thus, should a member of the team refuse to participate in the emergency operation, it would feel inappropriate to accuse him of doing a *wrong* thing.

Now, helping a person in distress might be seen as, in principle, a similar supererogatory act: it is good to do it (things are in moral equilibrium), but not wrong not to do it. The implication here bearing on (F2) would be the following: if helping a person in distress qualifies as a supererogatory act and supererogatory acts are non-obligatory (i.e. it is not wrong not to do them) then the moral equilibrium doesn`t get disrupted (i.e. things don`t get *unfair*). At the same time, it has been agreed that refusing to save a drowning child is wrong. This might suggest that *right-wrong* and *fair-unfair* are not interchangeable notions after all. And if they are not interchangeable, I cannot translate Rawls` `sense of justice`, which is meant to be about a capacity to detect what is and isn`t *fair*, into a capacity to distinguish *right* from *wrong*.

I am somewhat sceptical regarding the notion of a supererogatory act and I don`t think that there are any uncontroversially supererogatory acts. But here is not the place to discuss it. I will assume, for the sake of the argument, that the case of the emergency technicians dispatched to shut down a faulty nuclear reactor is a clear case of a supererogatory act. The case is, at the same time, importantly different from the case of a refusal to save a drowning child. The action of the emergency technicians involves sacrificing their lives, which is something that saving a drowning child doesn`t involve. Not only doesn`t saving a drowning child involve a sacrifice of one`s life, it doesn`t involve a sacrifice of any kind apart from, perhaps, something like getting your clothes wet or arriving late for a meeting. Peter Singer, in a similar context, proposes the following moral principle:

> [I]f it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it.[264]

Singer treats the principle as uncontroversial and so shall I. According to Singer, and common sense, a human life counts as something morally significant. Thus, sacrificing it violates the explicit condition of the principle.

Employing Singer`s principle will let us draw a clear separation line between the case of emergency technicians and the case of a drowning child. The technicians find themselves in a situation that involves sacrificing their lives. This violates the explicit condition of Singer`s principle. It follows that the technicians are not obliged to perform the action. The

[264] Peter Singer, `Famine, Affluence and Morality`, *Philosophy and Public Affairs*, 1, no. 1 (1972), pp.229-243 (p.231).

case of a drowning child, on the other hand, doesn`t involve sacrificing anything morally significant and, therefore, one is obligated to jump into the pond and save the child.

Now, there are two crucial points here: Firstly, and less importantly, the case of a drowning child clearly involves a moral obligation to do so, which means it would be wrong to refuse to do it. As such the case of a drowning child is not a case of a supererogatory act. Secondly, and more importantly, the fact that the supererogatory case of the emergency technicians doesn`t involve a moral obligation indicates that *right* and *wrong* are used in a *non-moral* sense here. The overwhelming majority of moral realists believe that moral facts are obliging.[265] Therefore most of my readers (remember: I have lost a moral non-realist above[266] when tying the relevance of the problem of free will for a moral theorist with the notion of *fairness*/*moral equilibrium*; which is unacceptable for a moral non-realist) will be happy to accept that if a situation has no obliging force then it is not open to descriptions in terms of *moral* rights and wrongs.[267]

Let me sum up the discussion of (F2). Rawls` notion of a sense of justice refers to a capacity to detect what is fair and unfair. In (F2), I modify the notion as referring to a capacity to distinguish right from wrong. This modification faces an objection. The objection deploys the concept of a supererogatory act: an act that is good to do but not wrong not to do. The possibility of a supererogatory act is meant to imply that the notions of *right* and *wrong* are not logical opposites. That is a problem for my modification in (F2) because the notions of *fair* and *unfair* are logical opposites. This would mean that the notions *right* – *wrong* and *fair* – *unfair* are not interchangeable and that the *sense of justice* as a capacity to detect whether a state of affairs is or isn`t *fair* cannot be modified into a *sense of justice* as a capacity to distinguish *right* from *wrong*. I employ a simple argument to counter the objection: The moral is obligating. Supererogatory acts are not obligating, that is, they are not *moral* acts. Therefore, *right* and *wrong* must, when used to describe a supererogatory act, be understood as *non-moral* notions. This fully neutralizes the objection. The *rights* and *wrongs* of a supererogatory act are not the *rights* and *wrongs* of a morally significant situation. Thus, that which applies to the former doesn`t necessarily apply to the latter.

[265] Notable exceptions are Michael Smith – see his *The Moral Problem* (Oxford: Basil Blackwell, 1994) – and Peter Railton – see his `Moral Realism`, *Philosophical Review*, 95, no. 2 (1986), pp.163–207.

[266] See Chapter 4; especially sections 4.5.3.3 to 4.7.

[267] The reader might have noticed that when discussing the scenario of the emergency technicians above, I used the expression `good` instead of `right`. Thus, I said that `it is a *good* thing to shut down the reactor` and not `it is a *right* thing to shut down the reactor`. Just like this, without a context, it, perhaps, feels ok to say both *good* and *right*. However, once we formulate it as related to an agent then it won`t feel so. Compare: `The emergency technicians did a (extremely) *good* thing when they, sacrificing their lives, shut down the reactor to save the nearby city` and `The emergency technicians did the *right* thing when they, sacrificing their lives, shut down the reactor to save the nearby city`. Also, notice that the very definition of a supererogatory act is given in terms of *good*-ness and not of *right*-ness. All this strongly suggests that supererogatory acts are not *morally* significant acts.

1.1.2.3 Within the Christian paradigm, the New Testament is the foundational text regarding the moral realm. It has been long recognized as a text that is correct about moral facts

There is something I need to get out of the way first. So far, I have been addressing *all* moral realists. My strategy has been to show how the commitment to moral realism, when combined with certain plausible assumptions, implies that a moral realist should take the moral message of the New Testament seriously. (F3), however, introduces a rather significant restriction. It places the argument within the confines of `the Christian paradigm`, which is a step that seems to come at the cost of losing all the *non-Christian* moral realists. There are three things I wish to say in response.

First, I concede that with (F3) I am losing a non-Christian moral realist. It is not a huge loss though. I write within the tradition of Anglo-Saxon analytic philosophy. Most of my readers will be Christians in the broad sense of having grown up within and having internalized Christian culture.

Second, even though I concede I have lost a non-Christian moral realist, I have done so only because keeping the argument all-inclusive with respect to different religions and cultures would require a rather lengthy argument and a deviation from the main argument - a deviation that I don`t wish to waste any time with here. I am, however, convinced that the argument could be further developed so as to be all-inclusive. Such a development would employ a broadly Millianistic strategy and argue that the relevant moral and religious texts of different religions refer – via differing descriptions - to the same moral facts.

Third, there are some significant differences regarding the accepted moral truths across different cultures and religions. Not uncommonly, the moral truths accepted by different cultures and religions will contradict each other. Thus, for instance, a Hindu man is committing no sin when using contraception while a traditional Jew is. This particular example of differing moral attitudes towards contraception is no minor issue. It is, ultimately, about whether it is *right* or *wrong* to interfere with the natural order of things to prevent a new human being to be born. The fact of contradicting moral truths (across different cultures and religions) poses a problem for a moral realist. The problem has a form of the following dilemma: either some of the cultures and/or religions got some of their (often fundamental) moral facts wrong or, somehow, the world is such that two moral claims can both be true and contradict each other. The latter is off-putting and unacceptable. The former, clearly, is not such a crazy position as the latter one, yet it is a rather unattractive one all the same. One of the implications of this position would be, for instance, that those cultures or religions that got some of their moral facts *wrong* (or just got *more* of them *wrong* compared to other cultures and religions) are morally *inferior* to those cultures and religions that got the facts *right*. Another problematic implication would relate to the status of our sense of justice. It is, as explained above, a capacity to distinguish *right* from *wrong*; i.e. the sense of justice is a capacity to get the moral facts right. The capacity is meant to be a *universal* human capacity. What would explain a culture or religion, that is, a large number of people (often hundreds of millions), getting some of their moral facts wrong for many centuries? How could their sense of justice have been collectively malfunctioning for so long?

My point here is this: True, having placed the discussion within the confines of the Christian paradigm, I have lost a non-Christian moral realist. However, this placement of the discussion within the confines of a religion is not something I impose onto my reader (a moral realist) because my argument requires it. It is, rather, something that follows from the problems implicit in moral realism itself. Whenever it comes to the discussion of specific moral claims, a moral realist has to choose one of the (empirically) available moral frameworks (has to *place* the discussion within one of them) to evaluate the claims because, unfortunately, there are more of them and they, to an extent, contradict each other.

Now, back to (F3): Within the Christian paradigm, the New Testament is the foundational text regarding the moral realm. It has been long *recognized* as a text that is *correct* about moral facts. (F3) consists of two related claims. Both are, I believe, fairly uncontroversial. Undeniably, the New Testament is the foundational text of Christianity. Within the paradigm of Christianity, the New Testament provides moral guidelines[268] for its followers. *Imitatio Christi* is the ultimate moral ideal. Similarly undeniable is the fact that the New Testament has been long *recognized* as a text that is *correct* about moral facts. By `recognized` I don`t mean a simple `acknowledgement` of a fact; by `recognized` I refer to a deeper kind of acquaintance with and acceptance of the moral truths that the New Testament contains. The depth of the acquaintance and acceptance ranges from a relatively shallow internalization of the truths that is common among secular Christians to a very deep, experiential kind of acquaintance found among devout Christians. For a devout Christian, the moral veracity of the New Testament is not just *cognized*, it is *experienced*. A secular Christian doesn`t, perhaps, have the same (direct) experiential connection to the moral facts of the New Testament, yet she will reliably *recognize* the truth of the facts if confronted with them. There is a simple way to test the plausibility of this last claim. The reader of mine who is a secular Christian (and chances are that such is the majority of my readers) can test their moral intuition (that is, deploy their sense of justice) against the moral narrative of the New Testament. More specifically, the secular reader is invited to *morally* challenge the actions of Christ as described in the New Testament. How does it sound to say something like `Christ was morally *wrong* to do…`? Can you see yourself sincerely endorsing some such

[268] The moral guidelines in the New Testament are mostly implicit in the sense of being the `morals` of a story. A Christian is obliged to make his life an *imitatio Christi* in the sense of living within the framework of those `morals`.

proposition? I don`t expect you can. Even a secular Christian will find it extremely hard, if not impossible, to imagine predicating a morally wrong action about Christ and *mean* it. In this sense, even a secular Christian *recognizes* the New Testament as a text that is *correct* about moral facts.

Now, what I am saying here could be explained away as being just a semantic issue and not a *perceptual* one (i.e. not one that is about the *sense* of justice). `Christ` as a symbol or concept refers to, among others, something like `the ultimate moral goodness`. Thus, predicating a morally *wrong* action about Christ yields a more or less straightforward contradiction; and it is a *recognition* of this contradiction – and not our acquaintance with and acceptance of moral truths in the New Testament – that makes it impossible to be serious (or not confused) when predicating a morally wrong action about Christ.

This is a valid point and one that, to a great extent, explains what`s going on when one tries to associate Christ with moral wrongness. This is, however, also a point that might obscure the fact that the issue goes somewhat deeper here. True, we cannot associate Christ with moral wrongness because `Christ` *means*, amongst other things, something like `the ultimate moral goodness`. However, and importantly, this meaning hasn`t been *stipulated*. `Christ` has come to mean this due to referring to someone who has been presented (in the New Testament stories) as the agent in actions that people, historically, recognized as morally deep and *right*.[269] Thus, at the core, we have the same fact here as the one at the core

[269] This last point might strike some of my readers as contradicting one of the crucial claims in the first argument. There I argue that a moral fact always presents itself to us together with an obliging force, and that this obliging force originates in God. Christ is God which should mean, following the logic of the first argument, that the reader recognizes the message in the New Testament as morally right (or true) only after he has already recognized Christ (the agent of the actions described there) as God. In the second argument, however, the direction of explanation gets reversed. There I seem to imply that Christ is recognized as God (as `God` because only God is `the ultimate moral goodness`) because he is the agent in stories all of which people have recognized as describing morally right actions. We have, admittedly, a bit of a contradiction here. I don`t intend to provide a successful resolution of it here, which is mainly because I don`t need to somehow combine both arguments (the first one and the second one) to be able to conclude that a moral realist has to take foundational religious texts seriously. Each of the arguments alone, once accepted, yields the conclusion. Thus, any contradiction between the two is non-consequential here. Still, let me, very briefly, gesture towards a strategy to resolve the contradiction. The problem is, roughly, this: It seems to be correct to say *both* that we take (someone called) Christ as God because he did things (and *only* things) that we recognize as morally right *and* that we recognize certain things as morally right only after having established they were done (or commanded) by Christ (who is God). Now, I believe it could be argued that the two recognitions come *together* and *gradually.* The first followers of Christ had some indicators that Christ *might* be of a divine essence (it had been prophesied that a Messiah would come; Christ performed a few miracles; He explicitly introduced himself as the Son of God, etc.). At the same time, he both preached and did things that *struck* (i.e. they were quasi-recognised by) his followers as *potentially* conveying deep and important moral truths. At the beginning all was extremely ambiguous. He could have been a false prophet. But gradually what he preached and how he acted worked together with who He claimed to be. The more what He said and did was recognized as morally right, the more His divinity was recognized and vice versa: the more His divinity was recognized, the more what He said and did was recognized as morally right. And, gradually, with this interrelated double recognition growing, the obliging force of the perceived moral truths too was growing.

of the second claim in (F3): With our (universally human) sense of justice we *recognize* (perceive) the correctness of moral truths contained in the New Testament.

1.1.2.4 The New Testament is correct about moral facts

(F4): The New Testament *is* correct about moral facts. I move from having a *recognition* of correctness of something, (F3), to that something *being* correct, (F4). However, it is not obvious that the move is valid. True, `to recognize` is a *factive* verb and as such it is normally taken to mean something like `to arrive at a *true* belief` (about what is the case). When you say, `Among the things scattered on the floor, John has recognized his wife`s car keys`, what you imply is that the keys John has identified as belonging to his wife indeed *are* his wife`s. The problem is that it might turn out, as it often does, that you haven`t *recognized* anything after all. Thus, it sounds perfectly normal to say something like: `I recognized Mary in the crowd but when I got closer, I realized I was wrong`.

The important bit to notice here is this: when a speaker uses a factive verb, she indicates a *commitment* to the truth of the embedded proposition. This *commitment* of a speaker to the truth of the embedded proposition doesn`t, however, necessarily translate into that proposition being actually true. If it did, we could just make propositions true by making such commitments. And that`s absurd.

Now, this possibility of *mis-recognizing* things then casts a shadow of doubt on the move from (F3) to (F4). Or so it seems. How serious an objection is this? Not very, I believe. It is just a version of a familiar sceptical challenge. It seems to be an inescapable predicament of our perceptual-cum-cognitive relation to the world that we can, and often do, hold wrong beliefs about the world. A sceptic likes to draw far-reaching and radical conclusions from this inescapable predicament. She will insist that we can`t really know what the world is like – or what is really the case out there in the world – beyond the veil of perception. Analogously, a sceptic will insist that you cannot know what the world is like beyond your *recognition* and therefore you cannot conclude anything about the state of the world from having a *recognition* of it.[270]

Of course, the sceptic is wrong here as she infers *too much* from the possibility (and/or actual occurrence) of a perceptual and/or cognitive error. The mere *possibility* (and/or actual occurrence) of a perceptual and/or cognitive error in no way implies that our perceptual-cum-cognitive relation to the world *actually* is such that we can`t and don`t know what the case *is* out there in the world. It is one thing to say that we sometimes (or often) err about the world and entirely another to say that error is *intrinsic* to our perceptual-cum-cognitive interaction with the world. The latter simply doesn`t follow from the former. Thus, when I

[270] Recall: `having a *recognition*` of something is understood here as something like `being *committed* to the truth of the propositional content of a perception` of something.

perceive a tree out there, I have an extremely good reason to believe that there really *is* a tree out there. I might be wrong as I might be hallucinating the tree. However, the default position is that perceiving the tree implies its existence, therefore that`s what I will believe until confronted with convincing reasons that I have, in fact, hallucinated the tree. Similarly, the default position regarding our sense of justice is that it is a perceptual capacity of ours that, on the whole, connects us *reliably* with the moral realm. We can be and often are mistaken about moral facts, but we have no reason to believe that error about the moral is *intrinsic* to our perceptual access to the moral domain.

I shall conclude that there is no serious reason to reject the move from (F3) to (F4).

1.1.3 An issue with recognizing moral truths

There is a little issue here that needs to be briefly addressed before we can move on to the actual discussion of the above quoted passage from the New Testament.

I have promised to `extract [philosophical] reasons` from the quoted passage. Now, one might wonder why, or in what sense, I need to do any *extracting* at all. In this chapter, I argue, ultimately, that the kind of holding others responsible that requires grounding in the free will assumption is a kind of performance which is morally *wrong*. I seek to find support for this claim in the text of the New Testament. At the same time, I argue that we have, more or less, straightforward access to the moral truths of the New Testament. So, in what sense do I need to process the text *theoretically* to get to its moral truths? Can`t I just refer to the relevant passage and let the reader *recognize/perceive* the truth(s) for themselves? In fact, doesn`t the very need to do any *extracting* imply that (F2) cannot be correct, i.e. that it cannot be correct that we have a sense of justice that gives us a (in principle reliable) access to moral truths? The answer to these questions is relatively uncomplicated.

We have the ability to recognize/perceive all those moral truths that the New Testament is *specific* about. These *specific* moral truths combine with other moral or non-moral truths, which yields new moral truths. These new moral truths become available to us to be recognized as moral truths only after the combining that yields the truths has taken place. Prior to the combining they are not recognizable/perceivable in the text. I will stick with the story of the stoning of an adulteress to illustrate how it works.

The *specifics* of the story are about *adultery* as a sin, *stoning* as a punishment and the moral standing of those who gathered to stone the poor woman. We read the story and our sense of justice recognizes what Christ did in the story as morally *right.* We picture the poor woman who succumbed to temptation about to be stoned by a group of men some of whom might have done the same or even worse. We immediately recognize the striking *unfairness* of the situation. It feels obviously unfair to have those other sinners stone the woman to death. The moral truth that we have a straight recognition of here is: it is *wrong/unfair* if a sinner stones an adulteress. Now, one might wonder whether the truth of the claim is preserved when, for instance, \`stoning\` gets substituted by \`whipping\`, or when an \`adulteress\` gets substituted by a \`thief\`. The story talks *specifically* about \`stoning\` and \`adultery\` and makes no mention of whipping or theft. Could it be that if it was a *thief* about to be *whipped* then the floggers\` moral standing wouldn\`t matter at all?

I think we have a good reason to believe that the story extends to other (probably all) sins and kinds of punishment. Due to limited space here, I will show only how to extend the story to apply to all sins, not just adultery (a similar strategy could be then used to extend the story to all punishments, not just whipping). In the story, Christ encourages those \`without sin […] to throw a stone at her\`. It is clear that He doesn\`t talk about a *specific* sin but about sinning generally. It is *not* that *only* those who haven\`t committed the sin of, for instance, adultery can throw the stone. This fact strongly suggests that it is not the specific nature of a given sin (adultery in our case) that compromises the moral standing but rather the shared property of all sins – their moral wrongness – that does the compromising.

An interesting and important point here is that this implies a (perhaps somewhat surprising) asymmetry. Arguably, adultery is (or at least was, at the time of the New Testament) a more *serious* sin then the sin of lying. Christ\`s encouragement doesn\`t discriminate among sins, therefore even those who committed the rather innocuous sin of lying are morally compromised to throw a stone even though their target is a (perhaps much) bigger sinner.

We can conclude two things from Christ\`s encouragement. First, He talks about moral standing and not about a particular sin. Second, He implies that any symmetry in moral standing between the punished and the one punishing is irrelevant. Now, against the background of these two conclusions, try to imagine that somehow the story is *only* about the sin of adultery and doesn\`t extend to other sins. If this was the case, then we would have to ask what exactly it is about *adultery* (that other sins lack) that calls for the punishing one to be of an uncompromised moral standing. It can\`t have anything to do with the *seriousness* of the sin of adultery and its relation to the level of corruption that the moral standing of the punishing one might have sustained because, as concluded above, it is not about symmetry: even the least serious sin counts as sufficiently corrupting. Thus, it would have to be something about the very *nature* of adultery (and adultery *only*) that calls for a totally uncompromised moral standing of anyone wishing to punish it. Trying to figure out how this could even begin to be approached borders on the unintelligible. Once we know that (a) it plays no role exactly which sins have compromised the standing of the punishing one and (b) it is not about symmetry between the sins of the punished and the punishing one, we are able to direct our attention to a *general* relation (and away from the specifics of a particular sin) and conclude that the story applies not only to adultery but other sins too.

To sum up, one of the moral truths directly available to us in one of the stories in the New Testament is something like this: it is *wrong/unfair* if a sinner stones an adulteress. However, once we correctly combine this particular moral truth with some other relevant aspects of the story, we are able to generalize and extend it to be not only about adultery but about other sins too. This is an example of a way of *extracting* moral truths from a religious text. And such extracting is in no tension with the fact that we – as beings equipped with a sense of justice – have (a kind of) *perceptual* access to moral truths. Now, back to the actual discussion of the story of the adulteress.

1.2 The story of an adulteress and the moral standing of man

For some time already, the assumption has been that my reader is a moral realist. To avoid certain unnecessary complications, I have put a (innocent) constraint on the assumed position of moral realism here: I address a (broadly) *Christian* moral realist. Now, if I am right to claim (F4), i.e. if I am right to claim that the New Testament is correct about moral facts, then the reader (a Christian moral realist) must treat it as a serious source of moral knowledge. Above, I have promised to extract from the New Testament story of an adulteress reasons to believe that being a human comes *essentially* with having a deeply compromised moral standing. The reader will have remembered that the idea motivating this promise is to show that theorizing about free will in the context of justifying our practice of holding others responsible is pointless because the practice is irremediably unjustifiable due to its involving, inescapably, a holder with fundamentally compromised moral standing. So, what is it in the New Testament story of an adulteress that suggests that being a human comes with essentially compromised moral standing?

At the end of the discussion of the second argument, I deal with an objection, which involves arguing that the moral truth of the New Testament story of an adulteress is generalizable as something like this: punishing a sinner is morally wrong if done by a person whose moral standing is compromised. This is a very plausible and natural understanding of the story. But how do we get from here to the claim that a fundamentally compromised moral standing is *essential* to being a human? What in the story and in the broader context of the New Testament could support such a radical claim? The key to answering those questions lies in that part of the story where Christ joins the others in refusing to throw a stone at the adulteress saying: "Neither do I condemn you. Go your way, and from now on do not sin again." Now, consider the following simple argument:

M1: Principle (P): punishing a sinner is morally wrong if done by a person whose moral standing is compromised.

M2: (P) applies to Christ: Christ`s moral standing is compromised

M3: Christ is an archetypal human. This implies that whatever limitations can be predicated about Him will apply to all men.

Conclusion:

M4: The moral standing of all men is essentially compromised.

1.2.1 Principle (P): punishing a sinner is morally wrong if done by a person whose moral standing is compromised

(M1) is a very natural and widely agreed upon reading of the story of an adulteress. Importantly, there doesn`t seem to be any reason to object to this reading. It seems sufficiently clear that that`s the moral message of the story. Also, it is in full accord with what our sense of justice tells us is morally right and wrong here. We have a well-established notion of *hypocrisy* which refers to a moral truth that is analogous to (M1).

1.2.2 (P) applies to Christ: Christ`s moral standing is compromised

Christ is confronted with a sinner in the context of punishing her. And, as we know, he *fails* to punish her. He doesn't say anything that would indicate that he sees any problem with the punishment itself. Neither does He doubt that the woman indeed has sinned: `[D]o not sin again`, He instructs her. More importantly, Christ explicitly ties the punishment to the moral standing of anyone getting ready to inflict the punishment: `Let anyone among you who is without sin be the first to throw a stone at her`. All this strongly suggests that (P) applies to Christ, i.e. that Christ himself has had His moral standing compromised.

1.2.3 Christ is an archetypal human. This implies that whatever limitations can be predicated about Him will apply to all men

(M3) consists of two related claims. The latter claim is an implication of the former one; an implication that is rather straightforward once the truth of the former claim is established and understood. So, is it correct to say that Christ is an archetypal human? And what does it even mean to say something like this?

Two things are predicated about Christ here: being a human and being archetypal. Regarding the first of the two, let me simply brush away any doubts about Christ being a human. In the New Testament, He refers to Himself as `the Son of man` (see for instance, Mathew 8:20, Mark 10:32-34, Luke 6:5). For centuries, the Christological perspective on this self-reference of Christ as the Son of man has been seen as a counterpart to that of Son of God and just as Son of God affirms the divinity of Jesus, Son of man affirms His humanity.

Historically, there were differing and briefly influential teachings about Christ`s essence that, in some way or other, denied His humanity (such as Docetism, Apollinarianism or Eutychianism) but they have been denounced as serious heresies and abandoned. Moreover, and importantly, a moral realist who accepts the New Testament as a serious source of moral knowledge should reject any doubts regarding Christ`s humanity. She should do so because if Christ wasn`t a human then His life, acts and choices are not *relevant* to us humans. This would imply that the New Testament is, at best, a serious source of *irrelevant* (for us) moral knowledge; an implication that is, for all practical purposes, indistinguishable from *denying* that the New Testament can teach us anything at all about the moral.

Now, what about the claim that Christ is an *archetypal* human? In what sense is this correct to claim? To be able to answer the questions, I need to be a bit more specific regarding the notion of archetype (and archetypal). Let me distinguish two readings of the notion here: a philosophically *light* one and a philosophically *heavy* one. The latter implies a philosophical theory while the former doesn`t. To say, `They live in an *archetypal* country village` is to say something like `They live in a *typical example* of a country village` or `They live in a country village that looks very much like what the country villages *originally* looked like`. There is, clearly, no philosophy going on here. To describe a country village as *archetypal* is just saying either that there are many other very similar country villages or/and that it is a well-preserved example of what country villages used to look like. That`s the philosophically *light* reading of the notion.

The philosophically *heavy* reading can be found in Plato`s dialogues.[271] Plato doesn`t use the word `archetype`, that is, he doesn`t use the Greek form of it **-** `arkhetupon`. Instead, he uses the term *eidos* which, nevertheless, captures precisely the *heavy* reading of the notion of archetype. In Plato`s metaphysics, *eidos* refers to the original forms or (non-physical) essences of all things, to those first `blueprints` of everything that comes to exist in our spatiotemporal world. The word *archetype* comes from the above mentioned `arkhetupon`, which is a compound word formed of `arkhe` - meaning *beginning, origin, first place* – and `tupos` - meaning *a model*. `Arkhetupon` then refers to *the first* or *original model* (of something), which is analogous to what Plato`s *eidos* refers to. The *heavy* (or Platonic) reading of *archetype*/*archetypal* then refers to the non-physical essence that individuates *kinds* of things that exist.

Now, I claim that if Christ is an *archetypal* human then it implies that whatever limitations can be predicated about Him will apply to all men. Clearly, this implication cannot be sustained on the *light* reading of *archetypal*. On the *light* reading, the claim that

[271] The locus classicus is Plato`s Allegory of the Cave in his *Republic*. See Plato, *Republic*, Benjamin Jowett (trans.) (Oxford: Clarendon Press, 1888), 514a-520a.

Christ is an archetypal human would mean something like: Christ is a good example of what *most* men look and behave like, or what *most* men used to look or behave like. The implication, however, is about *all* man. We need something much more robust, something that brings *necessity* into the picture. And here the *heavy* reading comes to mind. The *heavy* (or Platonic) reading of *archetype*/*archetypal* takes us to *essences*, that is to what is *necessarily* the case about *all* humans. The *heavy* reading gives us the implication we need.

At this point, two questions arise. First, does anything like an *archetype* or *essence* of human exist at all? And, second, is it correct to say that Christ is or exemplifies such an *archetype*? It is quite tempting to go full Carl Jung here and show, deploying the conceptual machinery of his psychological theory, how, as a matter of fact, the complex cluster of properties (and the way they relate to each other) associated with Christ perfectly matches with those that Jung calls (the archetype of) the Self. `The archetype of the Self` refers to something like the objective or collective essence of man, and Jung himself argues that `Christ exemplifies [it]`.[272] The crucial point here – and one that answers the first question - is that the overall logic of Jung`s theory allows him to plausibly claim that his concept of (the archetype of) the Self is an *empirical* concept and not (merely) a *metaphysical* one (as in the case of Plato`s *eidos*). And, of course, being able to say that (the archetype of) the Self is an *empirical* concept amounts to being able to say that the Self (i.e. the archetype or essence of human) does exist. Thus, within the framework of Jung`s theory, it would be possible and plausible to give an affirmative answer to both questions.

I will not take the Jungian path here though. I have two reasons for not taking it. First, I can`t afford to let (M3) depend on something so complicated as is Jung`s theory of archetypes (and his reasons for taking them as empirically real). The scope of the chapter simply wouldn`t accommodate the lengthy exposition of Jung`s theory that such dependence would require. Second, and more importantly, I believe that there is an easier and more straightforward way to convince the reader that that there is a (sufficiently robust) sense in which essences or archetypes can be said to exist and that Christ exemplifies one of them.

Above, I have established that my reader – a broadly Christian moral realist – must treat the New Testament (that is, the story of Christ) as a serious source of moral knowledge. Taking the story of Christ seriously means taking it as prescribing what *one* ought to do, or how *one* ought to live. But how does it happen that a story of a particular *individual* – Christ – extends its normative force to (the generalizing) *one*? In virtue of what is Christ`s story *universalizable* in this way? Let`s call that in virtue of which Christ`s story is *universalizable* simply (and rather unimaginatively) *x*. Now, for our purposes here, there is no need at all to

[272] Carl Gustav Jung, *Aion: Researches into the Phenomenology of the Self (Collected Works of C. G. Jung)*, Richard F. C. Hull (trans.) (New York: Routledge & Kegan Paul, 1959), § 70.

go into the metaphysics of *x*, i.e. there is no need to evoke Plato`s or Jung`s theories. It suffices to realize and accept that there must be some such *x* for the story of Christ to extend its normative force to *all* people. My reader – a broadly Christian moral realist – accepts the fact of the extension of the normative force and therefore must assume *x*. Clearly, the logic behind *x* sustains the implication that whatever limitations can be predicated about Christ will apply to *all* men; and that`s what was needed.

A few words about the implication in (M3). At this point it should be clear what is meant by saying that Christ is an *archetypal* human. And if this is clear than it should be equally clear that this description of Christ implies that whatever limitations can be predicated about Him will apply to all men. Notice, however, that the implication is correct only if we talk about *limitations*. If we drop this from the implication, we get an implausible version of it: (if Christ is an archetypal human then) whatever […] can be predicated about Christ will apply to all men. This version is implausible because Christ is not only the Son of man but also the Son of God, meaning He also exemplifies an essence (a divine one) that *transcends* the merely human one. Thus, there will be many properties attributable to Him that won`t be attributable to humans. The relation between the human and the divine in Christ is that of the limited and the unlimited. Hence the need to insert `limitations` into the implication.

1.2.4 The moral standing of all men is essentially compromised

(M3) says that whatever limitations can be predicated about Christ will apply to all men. (M2) says it can be predicated about Christ that His moral standing is compromised. (M4) follows straightforwardly from (M2) and (M3) once it is accepted that having one`s moral standing compromised counts as a *limitation.* The notion of *limitation* as used in (M3) means something like a `limitation to perfection`. Thus the relevant part of (M3) could be paraphrased as `whatever *imperfections* can be predicated about Christ […]` without affecting the logic of the argument. I don`t expect any objections to the claim that having one`s moral standing compromised is an *imperfection*.

1.2.5 An issue with the argument

There is an issue here regarding the argument as a whole that calls for a brief discussion. Those of my readers who are as much as casually acquainted with the theological views of St. Augustine will recall his influential and (in)famous doctrine of original sin. According to the doctrine, Adam`s sin in Eden has got *inherited* by all men. All men are *born* morally corrupted. As a result of Adam`s primal sin, man is *essentially* morally corrupted regardless of whether he has ever sinned as an individual. That sounds exactly like the conclusion in (M4). The reader might wonder then why, instead of going through the hassle of formulating a new argument, I haven`t just made use of St. Augustine`s doctrine of original sin. There are two reasons I haven`t done so.

First, St. Augustine`s arguments behind the doctrine are too theological to figure in a philosophical thesis. The arguments are too theological in the sense of being inextricably embedded in the conceptual matrix of the speculations of the Church Fathers. The Church Fathers were deeply religious thinkers and the intuitions grounding their concepts were fundamentally shaped by their strong faith. To those lacking this strong religious faith (which is majority of my readers, I suppose), the arguments of the Church Fathers (with St. Augustine being one of them) will feel bafflingly unconvincing. Have a look at one of St. Augustine`s arguments behind the doctrine. The argument[273] can be structured in, roughly, the following way:

i. The sacrament of baptism is crucial for salvation of one`s soul.
ii. There is no alternative to baptism when it comes to the salvation of one`s soul.
iii. If there was an alternative, then it would mean that Christ had died in vain.
iv. It is undisputed that infants – who couldn`t, yet, have sinned and who don`t even have a capacity to do so – *need* to be baptised.
v. Conclusion: man is *born* a sinner.

The argument might be valid, but its explicit and implicit assumptions are way too constrained by a prior commitment to a complex theological doctrine. To a reader who is without such a commitment, the argument has no force.

Second, St. Augustine`s (and other Church Fathers`) understanding of the notion of sin is importantly different from how it is commonly understood. For St. Augustine and other Church Fathers, a *sin* is fundamentally a relational concept. It refers to `a *culpable* misrelation to God and the world`.[274] According to St. Augustine, the misrelation to God consists in replacing love of God with self-love.[275] Thus, when St. Augustine argues that each man inherits the primal sin of Adam and as such is born sinful, what he means is that each man inherits a misrelation, or is born misrelated, to God. Now, being essentially sinful in the sense of being essentially misrelated to God is a state that has nothing to do with being essentially sinful in a sense that morally disqualifies one from blaming and/or punishing others. Test your intuition on the following scenario:

[273] The argument is discussed in more detail in Jesse Couenhoven, `St. Augustine`s Doctrine of Original Sin`, *Augustinian Studies*, 36, no. 2 (2005), pp.359-396 (pp.361-362).

[274] Couenhoven, `St. Augustine`s Doctrine of Original Sin`, p.360.

[275] Some such claim is implicit in St. Augustin`s debate with Pelagians. See his *On Grace and Free Will* (GLH Publishing, 2017), especially chapter 33.

Your friend has been caught cheating on his wife. The friend and his wife argue. She is hurt and angry, `You are disgusting! I feel deeply embarrassed to have married someone so weak, cowardly and callous!`. Your friend will have none of it, `Who are you to talk to me like this?! You are so *misrelated to God*! You have no right, you hypocrite!`.

Surely, it feels very odd to question one`s moral standing – to call one a hypocrite – on the grounds of one`s misrelation to God. It feels too weak a ground at best and just an irrelevant one at worst. This probably explains why nowhere in his writings does St. Augustine himself draw any conclusions regarding the moral status of a blamer/punisher from his doctrine of original sin. The doctrine simply doesn`t support any such conclusions. If it did, it would be impossible, for a thinker of St. Augustine`s stature, to miss it.

Bibliography

Anderson, A.R., and Belnap, N.D., Jr., `Tautological Entailments`, *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 13, no. 1/2 (1962), pp.9–24

Anselm, Saint, *Proslogion*, M.J. Charlesworth (trans.) (London: University of Notre Dam Press: 1979)

Aristotle, *The Metaphysics*, Hugh Lawson-Tancred (trans.) (London: Penguin Classics, 2004)

Augustin, Saint, *On Grace and Free Will* (GLH Publishing, 2017)

Barnes, J., (review of) J. Hintikka, `Time and Necessity: Studies in Aristotle`s Theory of Modality`, *The Journal of Hellenic Studies*, 97 (1977), pp.183-186

Bayne, T., and Montague, M., (eds), *Cognitive Phenomenology* (Oxford: Oxford University Press, 2011)

Bealer, G., `Intuition and the Autonomy of Philosophy`, in Michael R. DePaul and William Ramsey (eds), *Rethinking Intuition* (Oxford: Rowman & Littlefield Publishers, Inc., 1998), pp.201-239

Becker, O., *Untersuchungen über den Modalkalkül* (Meisenheim am Glan: Anton Hain, 1952)

Beebee, H., `Does Anything Hold the Universe Together?`, *Synthese*, 149, no. 3, *Metaphysics in Science* (Apr., 2006), pp.509-533

Bible (*New Revised Standard Version*), John 7:53 – 8:11

Bishop, R.C., and Atmanspacher, H., `The Causal Closure of Physics and Free Will`, in Robert Kane (ed.), *The Oxford Handbook of Free Will* (Oxford: Oxford University Press, 2011), pp.152-170

Carruthers, P., `Invertebrate concepts confront generality constraint (and win)`, in R. Lurz (ed.), *The Philosophy of Animal Minds* (Cambridge: Cambridge University Press, 2009), pp.89-107

Coates, D.J., and Tognazzini, N.A., (eds), *Blame: Its Nature and Norms* (New York: Oxford University Press, 2013)

Cooper, J.W., *Panentheism: The Other God of the Philosophers* (Grand Rapids: Baker Academic, 2006)

Couenhoven, J., `St. Augustine`s Doctrine of Original Sin`, *Augustinian Studies*, 36, no. 2 (2005), pp.359-396

D'Arcy, E., *Conscience and Its Right to Freedom* (New York: Sheed and Ward, 1961)

Davidson, D., *Subjective, Intersubjective, Objective* (Oxford: Oxford University Press, 2001)

Dennett, D., *Elbow room: the varieties of free will worth wanting* (Cambridge, Mass.: MIT Press: 1984)

Dickie, I., `The Generality of Particular Thought` *The Philosophical Quarterly*, 60, no. 240 (2010), pp.508-532

Double, R., *Metaphilosophy and Free Will* (Oxford: Oxford University Press, 1996)

Dummett, M., "Can an Effect Precede its Cause", *Proceedings of the Aristotelian Society*, 28 (Supplement) (1954)

Evans, G., *The Varieties of Reference* (Oxford: Oxford University Press, 1982)

Faye, J., "Causation, Reversibility, and the Direction of Time", in Jan Faye, Uwe Scheffler and Max Urchs (eds.), *Perspectives on Time* (Boston Studies in the Philosophy of Science, Vol. 189), (Dordrecht: Kluwer Academic Publisher, 1997) pp.237–266

Fisher, J.M., *The Metaphysics of Free Will: An Essay on Control* (Oxford: Basil Blackwell, 1994)

`Recent work on moral responsibility`, *Ethics*, 110, no.1 (1999), pp.93-139

`Frankfurt-Type Examples and Semicompatibilism: New Work`, in Robert Kane (ed.), *The Oxford Handbook of Free Will*, 2nd edn (Oxford: Oxford University Press, 2011), pp.243-265

Fischer, J.M., and Ravizza, M., *Responsibility and control: A theory of moral responsibility* (New York: Cambridge University Press, 1998)

Frankfurt, H.G., `Alternate Possibilities and Moral Responsibility`, *The Journal of Philosophy*, 66, no.23 (1969), pp.829–839

The Reasons of Love (Princeton: Princeton University Press, 2004)

Friedman, M., `How to Blame People Responsibly`, *The Journal of Value Inquiry*, 47, no. 3 (2013), pp.271-284

Glock, H-J., `Animals, thoughts and concepts`, *Synthese*, 123, No. 1 (2000), pp.35-64

`Can animals act for reasons?`, *Inquiry*, 52, no. 3 (2009), pp.232-254

Hacker, P.M.S., *Human Nature: The Categorical Framework* (Oxford: Blackwell, 2007)

Harris, S., *Free Will* (New York: Free Press, 2012)

Hill, T.E., Jr., *Dignity and Practical Reason in Kant`s Moral Theory* (Ithaca: Cornell University Press, 1992)

Hintikka, J., *Time and Necessity: Studies in Aristotle`s Theory of Modality*, in *The Journal of Hellenic Studies*, Vol. 97 (1977)

`Aristotle on the Realization of Possibilities in Time` in Simo Knuuttila (ed.), *Reforging The Great Chain of Being* (Dordrecht: Springer-Science+Business Media,B.V., 1981), pp 57-72

Honderich, T., *How Free Are You?* (Oxford: Oxford University Press, 1993)

Horgan, T., and Tienson, J.L., `The Intentionality of Phenomenology and the Phenomenology of Intentionality`, in David Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary readings* (Oxford: Oxford University Press, 2002), pp.520-533

Hume, D., *A Treatise of Human Nature*, Book III, Part I, Section I

Jackson, F., `The Transitivity of Entailment`, *The Philosophical Quarterly*, 20, no. 81 (1970), pp.385–388

James, W., `The Moral philosopher and the Moral Life`, *International Journal of Ethics*, 1, no. 3 (1891), pp. 330–354

Janik, A., and Toulmin, S., *Wittgenstein`s Vienna* (New York: Simon&Schuster, 1973)

Joyce, R., *The Myth of Morality* (Cambridge: Cambridge University Press, 2011)

Jung, C.G., *Aion: Researches into the Phenomenology of the Self (Collected Works of C. G. Jung)*, Richard F. C. Hull (trans.) (New York: Routledge & Kegan Paul, 1959)

Kane, R., *A Contemporary Introduction to Free Will* (New York: Oxford University Press, 2005)

Kant, I., *Critique of Pure Reason*, F. Max Müller (trans.) (Doubleday, Garden City, N.Y.: 1966)

Grounding for the Metaphysics of Morals, James W. Ellington (trans.) (Hacket, 1993 [1785])

Korsgaard, C., `Kant`s Formula of Universal Law`, *Pacific Philosophical Quaterly*, 66, no. 1-2 (1985), pp.24-47

The Sources of Normativity (Cambridge: CUP, 1996)

The Constitution of Agency (Oxford: OUP, 2008)

Leibniz, G., *"Reflections on the Common Concept of Justice"*, in Leroy Loemker, *Leibniz: Philosophical Papers and Letters* (Dordrecht: Kluwer, 1989 [1702(?)]), pp.561–573

Lenman, J., `Compatibilism and contractualism: The possibility of moral responsibility`, *Ethics*, 117, no.1 (2006), pp.7–31

Lewis, D., `Are we free to break the laws?`, *Theoria*, 47 (1981), 113-121

Lewy, C., `Entailment and Propositional Identity`, *Proceedings of the Aristotelian Society*, New Series, 64 (1963 – 1964), pp.107–122

Libet, B., `Do We Have Free Will?`, *Journal of Consciousness Studies*, 6, no. 8-9 (1999), pp.47-57

Lovejoy, A., *The Great Chain of Being* (Harvard University Press, 1936)

Maclachlan, D.L.C., `The Pure Hypothetical Syllogism and Entailment`, *The Philosophical Quarterly*, 20, no. 78 (1970), pp.26–40

Macnamara, C., `Holding others responsible`, *Philosophical Studies*, 152, no. 1 (2011), pp.81-102

Mawson, T.J., "The Euthyphro Dilemma", *Think*, 7, no. 20 (2008), pp.25–33

McDowell, J., *Mind and World* (Cambridge, MA: Harvard University Press, 1996)

McGinn, C., *Truth by Analysis: Games, Names, and Philosophy* (Oxford: OUP, 2011)

McKenna, M., `Robustness, control, and the demand for morally significant alternatives: Frankfurt examples with oodles of alternatives`, in David Widerker and M. McKenna (eds), *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (VT: Ashgate, 2003), pp.201-217

`Ultimacy and Sweet Jane`, in *Essays on free will and moral responsibility,* ed. Nick Trakakis and Daniel Cohen (New Castle upon Tyne, UK: Cambridge Scholars Publishing, 2008)

Mele, A., *Motivation and Agency* (Oxford: Oxford University Press, 2003)

Nagel, J., `Factive and non-factive mental state attribution`, *Mind and Language,* 32, no. 5 (2017), pp.525-544

Nagel, T., `Moral Luck` in Daniel Statman (ed), *Moral Luck* (New York: State University of New York Press, 1993) pp.57-71

Nietzsche, F., *Beyond Good and Evil*, Walter Kaufmann (trans.) (New York: Vintage, 1966)

Pettit, P., *A Theory of Freedom* (Oxford: Blackwell, 2001)

`Deliberation and Decision`, in Timothy O`Connor and Constantine Sandis (eds), *The Companion to the Philosophy of Action* (Wiley-Blackwell, 2011), pp.252-258

Piaget, J., *The Moral Judgement of the Child* (London: Kegan Paul, Trench, Trubner & Co. Ltd., 1932)

Plato, *Euthyphro*, Benjamin Jowett (trans.) (Adelaide: Adelaide University Press, 2014)

Plato, *Republic*, Benjamin Jowett (trans.) (Oxford: Clarendon Press, 1888)

Railton, P., `Moral Realism`, *Philosophical Review*, 95, no. 2 (1986), pp.163–207

Rawls, J., `The Sense of Justice`, *The Philosophical Review*, 72, no. 3 (1963), pp.281-305

Rescher, N., *Luck: The Brilliant Randomness of Everyday Life* (New York: Farrar, Straus and Giroux, 1995)

Richards, J.R., *Human Nature after Darwin* (New York: Routledge, 2000)

Rogers, K.A., *Anselm on Freedom* (Oxford: OUP, 2008)

Russell, B., `The Philosophy of Logical Atomism` in Bertrand Russell, *Logic and Knowledge. Essays 1901-1950*, ed. R.C.Marsh (Allen&Unwin, London: 1956)

Scanlon, T.M., *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press, 1998)

Searle, J., 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3 (1980), pp.417–457

Shepherd, J., `Deciding as Intentional Action: Control over Decisions`, *Australasian Journal of Philosophy*, 93, no. 2 (2015), pp.335-351

Singer, P., `Famine, Affluence and Morality`, *Philosophy and Public Affairs*, 1, no. 1 (1972), pp.229-243

Smart, J.J.C., "Free Will, Praise and Blame", *Mind*, 70, no. 279 (1961), pp.291–306

Smith, M., *The Moral Problem* (Oxford: Basil Blackwell, 1994)

`Does the Evaluative Supervene on the Natural?`, in Michael Smith, *Ethics and the A Priori*, (Cambridge: Cambridge University Press, 2004), pp.208–233

Sosa, E., `Minimal Intuition`, in Michael R. DePaul and William Ramsey (eds), *Rethinking Intuition* (Oxford: Rowman & Littlefield Publishers, Inc., 1998), pp.257-269

Strawson, G., *The Secret Connexion: Causation, Realism, and David Hume* (Oxford: Clarendon Press, 1989)

`The Impossibility of Ultimate Moral Responsibility` in Galen Strawson, *Real Materialism and Other Essays* (Oxford: Clarendon Press, 2008), pp.319-336

`Cognitive Phenomenology: Real Life` in Tim Bayne and Michelle Montague (eds), *Cognitive Phenomenology* (Oxford: Oxford University Press, 2011), pp.286-325

Strawson, P.F., `Freedom and Resentment`, in John M. Fisher and Mark Ravizza (eds), *Perspectives on moral responsibility* (Ithaca: Cornell University Press, 1993), pp.45-66

Stump, E., `Alternative Possibilities and Moral Responsibility: The Flicker of Freedom`, *The Journal of Ethics*, 3, no. 4, The Contributions of Harry G. Frankfurt to Moral Responsibility Theory (1999), pp.299-324

Swift, J., *The Writings of Jonathan Swift*, Robert A. Greenberg and William B. Piper (eds) (New York: W. W. Norton, 1973)

Swinburn, R., *The Coherence of Theism* (Oxford, Clarendon Press, 1993)

Taylor, R., *Ethics, Faith, and Reason* (Englewood Cliffs, N.J.: Prentice-Hall, 1985)

Thompson, M., *Life and Action* (Cambridge: Harvard University Press, 2008), part II

Timpe, K., *Free Will: Sourcehood and Its Alternatives* (Continuum, 2008)

Tonry, M., (ed.), *Why Punish? How Much? A Reader on Punishment* (Oxford: Oxford University Press, 2011)

Urmson, J.O., "Saints and Heroes", in Abraham I. Melden (ed.), *Essays in Moral Philosophy* (Seattle: University of Washington Press, 1958)

van Inwagen, P., *An essay on free will* (Oxford: Clarendon Press, 1983)

Vivhelin, K., `Freedom, Foreknowledge, and the Principle of Alternate Possibilities`, *Canadian Journal of Philosophy*, 30, no. 1 (2000), pp.1-23

`How to Think about the Free Will/Determinism Problem` in Joseph Keim Campbell, Michael O`Rourke and Matthew H. Slater (eds.), *Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science* (Cambridge, Massachusetts: MIT Press, 2011), pp. 313-340

Vogler, C., *Reasonably Vicious* (Cambridge: Harvard University Press, 2002)

Watson, G., `The Trouble with Psychopaths`, in R. Jay Wallace, Rahul Kumar, and Samuel Freeman (eds), *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon* (Oxford: OUP, 2011), pp.308-24

Williamson, T., `Conceptual Truth`, *Aristotelian Society Supplementary Volume*, 80, no. 1 (2006), pp.1-41

Wittgenstein, L., *Culture and Value*, ed. G.H. von Wright, trans. Peter Winch (Chicago: The University of Chicago Press, 1980)

Online sources:

Anders, C. J., 'Dog Given a Home to Die In ... But He Decides to Live Instead', Huffpost, (2016) <https://www.huffingtonpost.co.uk/entry/dog-near-death-lives-in-new-home_us_56951b9ee4b05b3245da6a3a?guccounter=1> [Accessed 23 February 2018]

"choose", Cambridge Dictionary (Cambridge University Press, 2019) <https://dictionary.cambridge.org/dictionary/english/choose?q=to+choose> [Accessed 16 September 2019]

Lamperti, J., *Does Capital Punishment Deter Murder? A Brief Look at the Evidence*, 2010 < https://math.dartmouth.edu/~lamperti/my%20DP%20paper,%20current%20edit.htm> [accessed 17 September 2019]

"N.T. Quoted from O.T. Parallel Passages - Study Resources", *Blue Letter Bible*, 2019 <https://www.blueletterbible.org/study/misc/quotes.cfm> [Accessed 25 September 2019]

PhilPapers, 'The Philpapers Surveys', ed. by PhilPapers (2014). <http://philpapers.org/surveys/results.pl?affil=Target+faculty&areas0=0&areas_max=1&grain=coarse> [accessed 21 March 2019]

Smith, S., `Scientists: The Future of oldest tree species on Earth in peril`, *AP News*, September 14, 2017 <https://apnews.com/776e453d15674f1e9eb20af289d6e46e/Scientists:-Future-of-oldest-tree-species-on-Earth-in-peril> [accessed 26 June 2020]