Acta Philosophica Turkuensia
Vol. 1

THE PARADOX OF FREEDOM

Determinism, Free Will and Responsibility

Ville V. Kokko

Acta Philosophica Turkuensia
Vol. 1

The paradox of freedom

Determinism, Free Will and Responsibility

Ville V. Kokko

UNIVERSITY OF TURKU

SERIES EDITORS:

Juha Räikkä
Valtteri Arstila

Philosophy unit, Department of Philosophy, Contemporary History and Political Science
University of Turku
FI-20014 Turku
Finland

ISSN 3087-5943
ISBN 978-952-02-0440-2 (PRINT)
ISBN 978-952-02-0441-9 (PDF)

Painosalama, Turku, Finland 2025

Written by
MSocSc Ville V. Kokko
University of Turku
Faculty of Social Sciences,
Department of Philosophy, Contemporary History and Political Science, Philosophy
Doctoral Program of Social and Political Sciences

Supervised by
Professor Emeritus Olli Koistinen
University of Turku
Docent Susanne Uusitalo
University of Turku and University of Oulu
Professor Joseph Almog
University of Turku, University of California, Los Angeles
Professor Valtteri Arstila
University of Turku

Reviewed by
Docent Aku Visala
University of Helsinki
Docent Matias Slavov
Tampere University

Opponent
Docent Aku Visala
University of Helsinki

Chairperson (custos)
Docent Susanne Uusitalo
University of Turku and University of Oulu

ABSTRACT

This dissertation examines the question of how we should understand free will and moral responsibility. The traditional approaches to the question have often concerned the compatibility of freedom and responsibility with *determinism* and *indeterminism*. Determinism implies that given the present, only one future is possible, which means that human choices, also, are "determined in advance" in some sense. Meanwhile, if a choice is indeterministic, it seems to be partly out of the agent's control. Thus, both options seem problematic for both freedom and responsibility. I start my study by examining this problem but then move beyond it to examine the connection between free will and moral responsibility before formulating my own proposition for how both concepts are best understood.

I begin my study of the relationship between (in)determinism and free will by making clear which definition of (in)determinism I use in this work. One key point is how these concepts are always tied to a given *level of description*; the universe being (in)deterministic at the bottom does not equate it to being so on some higher level, such as that of human choices. I go on to present my own carefully formulated version of what I call the *randomness argument*. This consists of two related claims. First, indeterminism in choices contradicts a person's control over the in a specific sense, insofar as it has any effect. Second, there is no reason to want free will to be indeterministic that is not reducible to a bare intuition that it must be. I examine numerous objections and theories that could counter this argument and show that none of them do.

Next, I go on to argue that we have mixed intuitions about whether free will should be indeterministic or deterministic. However, since I argued there is no reason to want indeterminism other than such intuitions, I also argue that we will do best to adopt a *compatibilist* view of free will, in which it is compatible with determinism. However, I go on to examine why it might be that we have incompatibilist intuitions in the first place to see what exactly we should understand free will as involving. I argue that most forms of determinism indeed contradict free will, and what emerges as a suitable criterion for free will is *reasons-responsiveness*.

After examining free will, I turn to the concept of moral responsibility. I examine the common assumptions made about how moral responsibility is connected to free will on the one hand and differential treatment such as punishment on the other. I argue that none of these connections can be justified by intuition. However, instead of rejecting them, I argue that they can be justified by building on my model of freedom as involving reasons-responsiveness. If the fundamental purpose of holding responsible is to guide behaviour, it makes sense to limit this to cases where the person is responsive to reasons. I also show that this model can explain and justify a number of things about how we actually hold people responsible.

In the conclusion, I show how my model presents a unified, reasoned basis for justifying the ethics of moral responsibility and their connection to free will. Free will is explained as the ability to act on one's best reasons, whereas responsibility is based on the ability to make a difference. Two important challenges we must face are that we are not fully free in this sense – but can become more so – and that there is no principle that can always tell us when to assign responsibility.

KEYWORDS: free will, freedom of will, moral responsibility, responsibility, determinism, indeterminism, compatibilism, incompatibilism, desert, punishment

TIIVISTELMÄ

Tämä väitöskirja tarkastelee kysymystä siitä, miten meidän tulisi ymmärtää vapaa tahto ja moraalinen vastuu. Perinteisesti on yleensä kysytty, sopivatko vapaus ja vastuu yhteen *determinismin* ja *indeterminismin* kanssa. Determinismistä seuraa, että jos nykyhetki on kiinnitetty, vain yksi tulevaisuus on mahdollinen. Tämä tarkoittaa, että myös ihmisten valinnat ovat ”etukäteen määrättyjä”. Jos taas valinta on indeterministinen, se näyttää olevan osittain hallitsematon. Täten molemmat vaihtoehdot näyttävät ongelmallisilta vapaudelle ja vastuulle. Aloitan tämän tutkimuksen tarkastelemalla tätä ongelmakenttää, mutta sitten siirryn siitä eteenpäin ja tutkin yhteyttä vapaan tahdon ja moraalisen vastuun välillä. Lopuksi muotoilen oman ehdotukseni sille, miten molemmat käsitteet tulisi ymmärtää.

Aloitan tutkimukseni (in)determinismin ja vapaan tahdon suhteesta selventämällä, mitä (in)determinismin määritelmää käytän tässä työssä. Yksi keskeinen havaintoni on se, että nämä käsitteet ovat aina sidottuja tiettyyn *kuvauksen tasoon*. Se, että maailma on (in)deterministinen syvimmällä tasolla ei tarkoita, että se olisi samanlainen jollakin toisella tasolla, kuten ihmisen valinnan tasolla. Jatkan tästä esittämällä huolellisesti muotoillun version argumentista, jota itse kutsun *satunnaisuusargumentiksi*. Se koostuu kahdesta toisiinsa liittyvästä väitteestä. Ensiksi indeterminismi valinnoissa on ristiriidassa valitsijan hallinnan kanssa tietyssä määrittelemässäni mielessä sikäli kuin sillä on mitään vaikutusta. Toiseksi ei ole mitään syytä haluta vapaan tahdon olevan indeterminististä, paitsi sellaisia, jotka palautuvat intuitioon siitä, että näin on oltava. Tarkastelen lukuisia vastaväitteitä ja teorioita, jotka voisivat kumota nämä väitteen, ja näytän, ettei yksikään niistä onnistu siinä.

Seuraavaksi argumentoin, että meillä on ristiriitaisia intuitioita sen suhteen, onko vapaa tahto indeterminististä vai determinististä. Koska ainoa syy haluta indeterminismiä ovat intuitiot, väitän, että on parasta omaksua *kompatibilistinen* näkemys tahdonvapaudesta, jonka mukaan determinismi on sen kanssa yhteensopivaa. Tästä jatkan kuitenkin tarkastelemalla sitä, mitä inkompatibilistiset intuitiomme paljastavat siitä, miten vapaa tahto tulisi ymmärtää. Argumentoin, että useimmat determinismin muodot ovatkin ristiriidassa tahdonvapauden kanssa, ja sopivaksi kriteerille tahdonvapaudelle paljastuu *perusteherkkyys*.

Tarkasteltuani tahdonvapautta siirryn käsittelemään moraalisen vastuun käsitettä. Tarkastelen yleisiä oletuksia siitä, miten vastuu liittyy yhtäältä tahdonvapauteen ja toisaalta ihmisten eriarvoiseen kohteluun, kuten rankaisemiseen. Väitän, että mitään näistä yhteyksistä ei ole riittävää perustella intuitioihin vetoamalla. En kuitenkaan hylkää niitä, vaan argumentoin, että ne voidaan oikeuttaa laajentamalla malliani, jossa vapaus liittyy perusteherkkyyteen. Jos vastuussa pitämisen perimmäinen tarkoitus on ohjata ihmisten käytöstä, on syytä rajoittaa tämä sellaisiin tapauksiin, joissa ihminen pystyy reagoimaan perusteisiin. Osoitan myös, että tämä malli pystyy selittämään ja oikeuttamaan monia puolia siitä, miten käytännössä pidämme ihmisiä vastuussa.

Johtopäätöksessäni osoitan, miten esittämäni malli antaa yhtenäisen, perustellun oikeutuksen moraalisen vastuun etiikalle sekä sen yhteydelle tahdonvapauteen. Tahdonvapaus selitetään kykynä toimia omien parhaiden syiden mukaan, kun taas vastuu perustuu mahdollisuuteen vaikuttaa asioihin. Kaksi tärkeää haastetta ihmiselle ovat, että emme ole täysin vapaita tässä mielessä – vaikka voimme tulla paremmiksi – ja se, että ei ole mitään lopullista periaatetta, joka voi aina kertoa, ketä pitää vastuussa.

ASIASANAT: vapaa tahto, tahdonvapaus, moraalinen vastuu, vastuu, determinismi, indeterminismi, kompatibilismi, inkompatibilismi, ansaitseminen, rankaiseminen

Acknowledgements

This dissertation has been a long time in the making, so it would be especially hard to thank everyone who has helped me along the way individually.

Due to their changing positions at the university, I went through a lot of supervisors on this one. Thanks to Olli Koistinen and Susanne Uusitalo for guiding me throughout the process; Joseph Almog for discussions and sparring ideas; and Valtteri Arstila for handling some of the administrative burden and always being helpful and giving advice.

Likewise, thanks to my pre-examiners Aku Visala and Matias Slavov, especially for constructive comments that I foresee playing a large role in my editing this text to be published as a new book for a larger audience later. Thanks also to Aku Visala for agreeing to be my opponent.

One of the greatest challenges in working on a dissertation project is the mundane one of balancing time and money. I was fortunate enough to have funding for parts of my research time, and thus, I am grateful to those who gave me that funding: the Philosophy unit and the Department of Philosophy, Political Science and Contemporary History at the University of Turku for my first and last grants during this time; TOP-Säätiö and Varsinais-Suomen Kulttuurirahasto; and especially to Suomen Kulttuurirahasto, who provided the longest grant period. I also thank the Finnish Doctoral Training Network in Philosophy and the Doctoral Program for Social and Behavioral Sciences at the University of Turku for travel grants to conferences that allowed me to expand my horizons and make useful connections.

For ideas, support, conversations, and more, thanks to all my teachers, colleagues and students (I even had a few of the latter) at the Philosophy unit, and my friends at the Association for Doctoral Researchers of Social and Behavioral Sciences at the University of Turku. I also want to thank my parents Pirkko and Kai and sister Vilma for always supporting me.

Finally, to Jasmiini, who has taught me more than anyone else.

Turku, October 2025

Ville V. Kokko

Table of Contents

Figures

1 Introduction

> The modern philosophical debate about the freedom of the will, which seems to have begun by an exchange between Hobbes and Bishop Bramhall, has long since degenerated into a dialogue of the deaf; and nothing is to be gained by joining it.
> -Alan Donagal.[1]

If you saw the title of this work, you will guess that the above discouraging quote is something I intend to prove wrong. Yet, it also has a point. The debate about whether free will can coexist with determinism or not – between compatibilism and incompatibilism – seems to have been stuck in place for a long time now. It is my hope in this work to be able to wiggle at least some part of it enough to get it moving again. More than that, I want to take the discussion beyond the old questions about determinism which, it will turn out, is not necessarily the relevant question at all. When considering how free we are, we do not get the answers by scrutinising whether elemental particles are determined or not.

Those who do not actually know the philosophy of free will frequently pose the question about it thus: "Do we have free will, or is everything determined in advance?" This sounds as if it assumes that determinism is the opposite of freedom. However, the debate in philosophy has actually been going on around whether free will *is* compatible with determinism. Further, compatibilism seems to be the more common, even majority position.[2]

---

[1]Donagan 1987, p. 174.

[2]*The 2020 PhilPapers Survey* 2020. More precise URL: <https://survey2020.philpeople.org/survey/results/4838>. The percentage is 59.16% for "accept or lean towards: compatibilism," with numbers of respondents accepting

Of course, since this is a dialogue of the deaf, many philosophers, too, still take it for granted that this is the question that should be asked. They should certainly benefit from reading this dissertation as well.

Another thing that almost everyone takes for granted is that freedom of the will is related to moral responsibility. Besides free will and its relationship to determinism, my other topic here is to explain moral responsibility and its relationship to free will.

In this introductory chapter, I start by looking at the basic problems about free will and responsibility, provide an overview of the argument in the rest of this work, and look at a few other questions that concern the whole work, its themes, and how I will discuss and express things in this work.

1.1 The compatibility problems

The relationship between free will and determinism is problematic in two ways. First, brief examination seems to indicate that free will is not compatible with determinism. Second, a closer look seems to show that free will is not compatible with indeterminism either.

1.1.1 Freedom vs. determinism: The compatibility question

Thinking about the possibility that everything in the world is determined by some kind of natural laws easily leads to the thought that there is something there that is incompatible with free will. There are multiple related threads to this line of thought. If everything is determined by natural law, does that mean your choices and actions, too, are determined by something outside of you? If you are made up of matter that follows whatever laws govern the reactions of matter, does that mean you cannot make choices? Does it mean you cannot be guided by reasons instead of physical causes?

or leaning towards the two other largest categories far behind: libertarianism 18.83%, no free will 11.21%.

Does it mean that your conscious thoughts do not guide you? What about the sense that, in a free choice, you could have done otherwise? Determinism certainly seems to make that impossible in at least one obvious sense, in that only one future is possible.

Even if one does not ultimately conclude that determinism really makes free will impossible, the question of free will usually involves tackling the question of what determinism implies for it. That is the case in this work as well. Instead of "determinism or free will?" the real question here is what Robert Kane call *the compatibility question*:[3] is free will compatible with determinism or not?

1.1.2 Freedom vs. indeterminism: The intelligibility question

The notion that free will contradicts determinism has some intuitive support, easily leading to the idea that free will implies *in*determinism instead. The most obvious problem with this arises when we try to articulate just what this would mean, because on a closer look, it seems that freedom is not compatible with indeterminism either.

Suppose you are making a choice right now, and it is undetermined. If nothing determines what the choice is going to be, though, how are you in control of it? If any out of several options can come to be selected, you cannot be sure the one you want comes to be chosen. We may have worried your reasons do not determine your choice if laws of nature do, but if the choice is truly indeterministic, your reasons cannot fully determine it either, meaning you might act against them. The choice, it seems, does not really lie in you – it might as well be a random event coming from the outside. And if we think new should be able to do otherwise, still we do not think it should always be possible that we might choose any option even if the others are better. Does this not make our choices something we have to fear, in that we can never be sure what they will be?

Thus, we also need to ask whether indeterminism is compatible with free will. Kane speaks of the *intelligibility question*, as the question concerns whether it is

[3]Kane 2002b, p. 9.

intelligible to say that freedom is compatible with indeterminism, or in other words whether an intelligible version of freedom under indeterminism can be formulated.[4] In their own ways, both questions are about compatibility between freedom and something else, as well as about the intelligibility of freedom in conjunction with something else, but I will use the terms the way in which they have been introduced here. I do find it somewhat more appropriate to refer to compatibility when speaking of freedom and determinism and about intelligibility when speaking of indeterministic freedom. It turns out that free will under determinism is easier to formulate intelligibly, but that does not guarantee that it will seem that real compatibility has been achieved.[5]

1.1.3 Vanishing responsibility

It is usually assumed that free will implies responsibility and the other way around, so questioning the possibility of freedom leads to questioning the possibility of responsibility. However, responsibility brings its own interesting flavour to the question. In this subsection, I present a perspective based on responsibility that seems to intuitively threaten agency under determinism, but also indeterminism.[6]

The strange shift in perspective that easily threatens the sense that anyone is responsible happens when we start viewing a person's acts as their own and made by the person at the moment when they happen, and then we shift back to see what led to those acts.

To introduce this topic, I borrow an extreme case used for roughly the same purpose by Gary Watson.[7] The case is the true story of Robert Harris, who was

[4]Kane 2002b, p. 18.

[5]For a different but related "intelligibility problem", see Pink, 2011, p. 355.

[6]For similar perspectives, see e.g. Slattery 2014, pp. 242–243, Honderich 2002, p. 102, Donagan 1987, p. 166, Harris 2013, p. 21.

[7]Watson 2002, pp. 131–143. For a more concise example of this case being used to make the same kind of point, see also Fischer & Ravizza 1993, pp. 1–4.

convicted to death over two cold-blooded murders he committed in 1978. To start with, Harris and his brother merely intended to steal two young men's car to use it in a bank robbery – certainly not the most innocent motive, but it seems so compared to what happened next. After they had hijacked the car at gunpoint and had verbally agreed with the youths that they could go free, Harris suddenly shot one of them in the back, chased the other down and shot him several times, and then came back to finish the first one off by placing the gun in his mouth and firing it. Not only did he kill his victims for little or no reason, he light-heartedly joked about it and casually ate the lunch they had left behind later – while his brother provided a convenient comparison of a more normal person's (even criminal's) reaction by running off to the bathroom to be sick at the thought. This extremely callous and utterly remorseless attitude makes Harris seem even more grievously responsible than if he had merely done the deed. It makes him sound like the strongest and most obvious example of someone who deserves to be punished.[8]

The twist comes when we hear about Robert Harris's inhumane childhood and what pushed him to become such a monster. He was born prematurely thanks to his father abusing his mother (possibly trying to cause a miscarriage), he had fetal alcohol syndrome, and he was constantly and severely abused by both his parents, leaving him with an urge to hurt and destroy everything, including himself.[9] Particularly pointed were words quoted from his sister: "I saw every grain of sweetness, pity and goodness in him destroyed. It was a long and ugly journey before he reached that point."[10] Such a perspective easily makes us question whether someone so shaped by nothing but abuse should be considered responsible at all – it is a completely different view than we get from merely considering his monstrous

[8]Watson 1993, pp. 130–134.

[9]Watson 1993, pp. 134–137.

[10]Cited here from Watson 1993, p. 134.

behaviour at the end of the road, seen only as free choices flowing from his character.

Even in ordinary cases, we can easily make responsibility vanish by analysing back far enough. It seems that a person is usually responsible for their own achievements or transgressions... but it is easy to make it seem otherwise just by looking at things differently. Suppose Alice is a miserable criminal and Bob is a successful philanthropist. Are they responsible for what they are – and what they do? Does Bob deserve higher esteem? What if he came from a wealthier and happier family than Alice? He would presumably have had a better chance to be good and successful than she. All right, Alice and Bob are not responsible for their background, and perhaps not for some things that are affected by it. Yet, what if Bob did have an equally miserably background? Then we would have a difference between Alice, who fell to the disadvantages of her background, and Bob, who is surely responsible and praiseworthy for not going the same way. In that case, what made the difference between their fates? Was Bob more intelligent, more strong-willed, more empathetic perhaps? Praiseworthy qualities, by normal thinking... but surely Bob is not responsible for being born with those properties any more than Alice is for being born without them. If, instead, Bob developed those qualities (while Alice did not), then maybe this was because there was something about his inborn abilities or environment that was different than for her – and again, neither of the two was responsible for that, since they had no control over it. If, on the other hand, there was some kind of chance, some kind of indeterminacy, causing the difference between Alice and Bob, that would *also* be something they could not control. It seems that it was all a matter of luck – and if they, or anyone, is praised or blamed morally, that is just moral luck.

Thus, just as both determinism and indeterminism seem to contradict free will, they seem to contradict responsibility also. Either there is a reason for everything that we do that is ultimately outside of us, or there is no reason for something we do, and “no reason” can hardly be our responsibility either. Look far back enough, and responsibility always seems to vanish with logical inevitability.

1.2 "The will"?

I speak of "free will" or "freedom of the will" as being one of the main' topics of this work. Looking at the form of those expressions, it would be easy to think that there must be some thing called the "will" that is free or not. Indeed, there has existed a notion of a "will" as a human "faculty". To quote *The Shorter Routledge Encyclopedia of Philosophy*:

> As traditionally conceived, the will is the faculty of choice or decision, by which we determine which actions we shall perform. As a faculty of decision, the will is naturally seen as the point at which we exercise our freedom of action – our control of how we act. It is within our control or up to us which actions we perform only because we have a capacity to decide which actions we shall perform, and it is up to us which such decisions we take. We exercise our freedom of action through freely taken decisions about how we shall act.
>
> From late antiquity onwards, many philosophers took this traditional conception of the will very seriously, and developed it as part of a general theory of specifically human action. … From the sixteenth century on, this conception of the will and its role in human action met with increasing scepticism.[11]

This work will simply use "free will" or "freedom of (the) will" as a whole as standing for that phenomenon, or quality of agents, that has been generally talked about under those names. It is not that we have a "will" that is free or not (or that we either have a "will" or not), it is that we either have "free will" or not, or maybe various degrees of it. Just what that means, in turn, remains to be determined; see sections 1.5 and 3.1.

That said, it is possible to interpret the "will" in "free will" in such a way that works in the context of this study as well. Within the normally used meanings of "will", a person's will is what that person wants, and we could say that free will is

[11]Pink 2005.

about the person's "wantings" to be free in some sense that remains to be specified. We could also go further and say that the person's will refers to that desire that actually moves the person to action, as opposed to every desire.[12] We could also say that "the will" is whatever process or capacity in us corresponds to the supposed faculty in that that process or capacity in us that makes choices. All of these options are so vague that they are possibly correct (to describe what is meant by "free will" in this work) in the sense that they are possibly not wrong, but they are also so vague they are not very helpful. Thus, the reader might do best to forget about "the will" by itself.[13]

If I was coining a new, accurate term for what I (and, probably, most others) are talking about when using the expression "free will", I might use the words "inner freedom".

1.3 Different answers and theories

Different answers given to the problems and questions posed above lead to different positions with respect to determinism and freedom. I will present these terms in the form in which I will be using them in this dissertation; the definitions I give are in accordance with common usage but not completely universal. As always, after I have introduced the terms, if I use them without further explanation below, I will always mean for them to be understood in the sense that I introduced.

[12]This idea is inspired by Frankfurt 1971, p. 8.

[13]Slattery 2014, chapter 21 has a useful account and interpretation of "will" or "willing", as long as you remember that he insists on using words with a particular meaning after a brief argument, so it is only about that kind of "willing" – and he also uses "free" to imply *undetermined*.

- *Compatibilism* is the idea that free will and determinism[14] are compatible.[15]
- *Soft determinism* is the idea that determinism is true of the world and compatibilism is true.
- *Incompatibilism* is the idea that free will and determinism are incompatible.
- *Hard determinism* is the idea that determinism is true of the world and incompatibilism is true, making free will impossible.
- *Libertarianism* is the idea that incompatibilism is true and there exists free will. I will not say, at this point, that it means there is indeterminism in our decisions, since libertarians sometimes deny this implication (see e.g. section 3.4.2, below). This meaning is unrelated to political "libertarianism".[16]
- *Hard incompatibilism* is the idea that free will is compatible with neither determinism nor indeterminism, and this means it is impossible.

---

[14]For now, I am ignoring the question of what level of description the determinism in question applies on (see 11.3 below), but the answer to that question for these definitions would be "On the ultimate level." Notice that this means that I am not using these terms as relative to different levels, so I cannot say, for example, that someone is a "compatibilist" on the ultimate level but a "libertarian" on a higher level (as Christian List would say about himself, see 3.4.3). More precisely, I should say I can do this when making it explicit, but then the terms are to be understood differently due to being explicitly modified by level, as in "otherwise this term as I have defined it except now modified by level like this." It would make sense to define these terms as relative to level, but the discussion about compatibility almost always assumes determinism on the ultimate level as being the main kind at issue.

[15]There is, as far as I know, no term for the idea that sometimes accompanies compatibilism that freedom *requires* determinism. Conversely, Manuel Vargas briefly calls the idea that determinism is compatible with *both* determinism and indeterminism "supercompatibilism", without introducing it as a proper term (2013, p. 13).

[16]That said, Yuval Noah Harari makes a somewhat convincing claim, though without proper evidence, that free will in a kind of libertarian sense is related to a "liberal" conception of the world in such a sense of "liberalism" that this amounts to at least close to saying free will libertarianism, while not meaning the same thing, is related to political "libertarianism". (Harari 2017, pp. 327–328; cf. also Harari 2017, pp. 299–301.) See also section 9.6.1 in the present work for a little more on this. I agree this at least may be true. When I say above that the two are unrelated, I am speaking in terms of what their definitions are.

Though all of the definitions above mention only free will, they could also be applied to responsibility, e.g. "incompatibilism with respect to responsibility". As mentioned above, this is often assumed to be equivalent or coextensive.[17]

There are also some terms used for more specialised positions that are not as important to remember while reading this work, but which I will mention for the sake of completeness:

- *Semicompatibilism:* John Martin Fischer's and Mark Ravizza's position according to which responsibility is compatible with determinism even if freedom is not.[18]
- *Partial compatibilism:* David Peroutka's view that free will and moral responsibility are compatible with determinism only in cases where the agent is psychologically compelled to do the right thing; if the agent does something wrong instead, that cannot be determined and free.[19] (For more on this, see 3.2.3.)
- *Compatibilist libertarianism:* Christian List's term for his position in which free will exists and is compatible with determinism on the ultimate level (see 10.6 in the present work) but requires indeterminism on some higher, psychological level (see 3.4.3). Not a form of libertarianism in my terms, merely bearing some resemblance to it.[20]
- *Attitudinism:* Ted Honderich's[21] term for his view that rather than either incompatibilism or compatibilism being wholly true in the sense of

[17]For a somewhat more detailed classification of basic positions, see Ofstad 1967, pp. 181–182.

[18]See Fischer & Ravizza, 2000 and e.g. Fischer, 2011.

[19]Peroutka 2022.

[20]See List 2019b, p. 12 for some of List's inspirations and similar positions.

[21]Honderich 2002, e.g. p. 154.

describing the only conception of free will, we have two different conceptions of freedom, one compatibilist and one incompatibilist. I agree with the core of this thesis in this work (especially 4.2.1), even though I choose to affirm the label of compatibilism because of further considerations (4.3), and I disagree with some of what Honderich says in 3.4.10.

1.4 How to ask and answer ultimate questions, part 1

The question of human freedom and responsibility is one that is tied intimately to how we see ourselves and our place in the world, and to our acting in the world and how we see that acting. True, one might not be interested in the technicalities of such questions – or one might only be intellectually interested in the philosophical technicalities. Nevertheless, once one sees the connection to the moral and otherwise important aspects of human life and its meaning, questions of freedom and responsibility become very sensitive. They also become strange in a way; something might be shown to be perfectly logical may still seem inadequate in the face of such important questions and such important aspects of our lives. One of the themes I am exploring in this dissertation concerns just this strangeness of questions with deep spiritual significance (spiritual in the sense of having to do with the meaning of life). That will especially be the theme in various parts of chapter 9.

In this section, I want to explain my considered opinion about how deep questions such as that of freedom and responsibility should be treated in a philosophical inquiry. I will lay out the principles now in a more abstract sense. After I have presented more of my arguments, I will return to the principles again and show how they have applied, in section 9.1. I am hoping this will help understand some of the potentially counterintuitive points I give arguments for.

Philosophy, especially analytic philosophy, is all about asking those questions that normally remain unanswered. The answers it gives should also be based on sound, explicit arguments. I will follow these principles in this work, but I will also

do something more. I will not make any conclusions based on anything but good argumentation, but since I doubt this will be enough due to the strangeness of deep questions, I will also endeavour to *make my point* and *illustrate* it in other ways.

If asking all the questions previously left unasked is the task of philosophy, then a very common sin of philosophers is not actually doing this. As will be shown in the forthcoming, it is all too common for philosophers to leave something unanalysed even within their analyses. This may be made explicit, but it can also be a question of building an analysis that fails to explain what it should because it already assumes it. For example, a model of how a (free) person makes decisions might contain a decision-making part that essentially acts as a person making decisions – in an unexplained way. (See 10.7 below.)

A closely related potentially bad habit of philosophers is taking too many things for granted. For example, I am going to ask (in chapters 7 and 8) why we think that a person is responsible for their acts under conditions such as freedom, and why responsibility can mean that it is permissible to harm that person in the name of punishment. In fact, I am going to ask *whether* this is really the case. Meanwhile, some of the philosophers whose views I discuss take these things as a given and build their theories around them (see especially chapter 6).

It is true that one cannot argue for every assumption that one uses in every context, nor can one analyse everything. It is not automatically wrong to leave unanalysed or take for granted. I do think these choices are made too often, and sometimes unknowingly. My argument here will be that in the question of free will and responsibility, it can be seen that these things have been done where they should not. In addition, I will show that I can ask and answer some such questions right here.

Taking some basic assumptions for granted without questioning them may be considered justified based on an appeal to intuition. This is discussed in more detail below in 6.3.1 and to some extent 4.3.3. I freely admit that it is a strong, commonly

shared intuition that some free acts make one deserving of punishment. As will be seen below, the content of this intuition has been considered a basic moral fact. I fully intend to take into account this and other intuitions about freedom and the related moral questions. However, I will be taking into account *the fact(s) of these intuitions*. We have these intuitions, they are part of our psychological and moral framework, and that is relevant. Fine. But why should we consider that our intuitions reveal the truth about something else? Why should we believe our intuitions indicate truth? They are, above all, psychological phenomena. The fact that I have an intuition is just the fact that I have a feeling that things are in some way, and (this is also part of the definition of intuition) the reasons for my having this feeling are not transparent to me. True, some of our deepest and most obvious beliefs are nothing but intuitions. True, the moral intuitions about freedom and responsibility may even be among these. Yet it is not the case, at least not in this case, that we are unable to examine these intuitions further. We can question them; we can demand arguments for them; we can see whether they can be explained psychologically; we can see whether we have reason to go on believing in them or reason to abandon them.

The above describes what might be loosely called the analytical side of my approach. In giving explanations and making arguments, we should analyse everything as well as we can, and likewise argue for everything we can argue for. This way, we can prove and explain things in as strong a sense as possible. Yet, as I said above, it may seem that this is not enough. I would say that it is not, although not in the sense that some (such as the opponents of compatibilism about free will and determinism) might claim.

There is one point of view that is certainly missing above. That is the pragmatic or practical aspect – and not just in a prosaic sense, but the existential level, the question of how we should live and see our lives. So, for example, in my analysis of the relationship between determinism and free will, I will first argue for what options are even possible, but after that, I will ask which possible option we should choose. It turns out intuitions about free will could be interpreted in more

than one coherent way, and the intuitions themselves cannot give a final answer as to which option is right. There are in fact contradictory intuitions. At this point it makes sense to ask: How should we readjust our concept of freedom so that it makes sense and serves the purpose it was supposed to have? (Section 4.3.3.) I will do the same with the concepts of desert and responsibility later on. (Chapter 8.) When doing this, I am taking into account the constraints facts and reasoning set to what is true and possible. That is not all, though. Such a perspective makes it possible to take into account that which is truly important about these concepts for us in our human existence. It respects the depth of the questions being answered – at least in some sense.

Yet, with deep questions, there is something more still. I could put forward arguments based on the above kinds of reasoning and be satisfied that anyone disagreeing on some kind of emotional grounds, without proper arguments, can be dismissed. However, if people are left with questions about how they could possibly understand what I present as being meaningful, there is some part of my job that I have failed to do. In order to do this last part, I will present thought experiments illustrating why there is reason to think the way I am proposing (e.g. 5.2, 6.3.1), and also describe how I see such things as our place in nature in a way that fits my theory (9.1.2). At these places, what I am doing is much more about illustrating and making my point than proving it; after all, as proofs, these things would be little more than appeals to intuition. Yet, I think this is a very important way of communicating the ideas when they are both hard to understand and concerned with the big questions in life.

1.5 Notes on notation and definitions

Words, terms, and concepts; italics and quotation marks

In careful philosophical examination, it is potentially important to know when you are talking about a word as in a series of sounds or letters that could be associated

with different meanings or none; when you are talking about a word as associated with a particular meaning; and when you are talking about the thing referred to by the word. The word "bear" has four letters; a bear is an animal and does not contain letters. It is also important to know what sense you are using a word in, and that you are using the word in the same sense every time you use it in a way such that it makes a difference.

To this end, I use a policy of denoting *words* (mere strings of letters), what I call here *terms* (words seen as being attached to a particular meaning) and *things* in different ways. Mere words will be in quotation marks, as with the "bear" above. Terms will be in italics, as with *words*, *terms*, and *things* above. Words denoting things will be used without either.

When trying to do this, it quickly becomes clear that it is difficult to make these distinctions with everything. It seems that the way we use words is so flexible that these three categories do not cover all the ways. Nevertheless, I apply this system as best I can. In addition, I also use italics when introducing new terms, and for emphasis, and quotation marks to indicate quotations and what might be termed "so-calledness". I believe that what I end up with is still less ambiguous than if I did not try to make the word–term–thing distinction clear in this way.

Since I write this text in natural language, I will unavoidably be using words all the time without giving them an exact definition or having one in mind. However, I try to be careful to give definitions to philosophically relevant key terms. There is a glossary of terms in the end (Appendix D), and I endeavour to use those words or terms always and only in the sense given there when I use them as words without quotation marks, italics or further explanation.

The key terms *free will* (or *freedom of will*, *freedom of the will*) and *responsibility* will be an exception in that I will use them liberally without having a specific definition in mind to start with. The reason for this is that looking at the discussion and the debate about them as I do here is largely a question of examining different definitions and seeing which ones should, normatively speaking, be

adopted. Thus, it makes no sense to commit to a definition from the start.[22] (See also 3.1.)

Cross-references

I frequently refer to other parts of this work for more information on some particular topic, giving section numbers. Note that numbers starting with 10 refer to Appendix A, and similarly for Appendix B being "11".

1.6 Overview

This work starts from some very basic concepts, goes on to explore free will, then moves on to responsibility, and finally comes to a conclusion uniting the two.

Chapter 2: A Brief Introduction to the Debate about Free Will and Determinism introduces the debate, as the name promises, as well as some useful concepts to use in it.

Before I can really discuss questions about determinism, free will and moral responsibility, I need to establish some basic concepts and tools that I use in the discussion. Thus, I start out by defining and explaining the concept of levels of description and giving my definition of determinism.

Levels of description are an aspect of all our talk about the world, something that once seen cannot be unseen and that comes up everywhere. In short, we can speak of the world in different sets of terms, all capable of describing the world truly though only to various different levels of approximation. Speaking of things happening on the ultimate physical level is very different from describing mental events – even if those mental events are emergent from those same physical things. When talking about determinism in particular, we need to know at which level of

[22]Though this is not something I have occasion to dig into more deeply, what I am doing with these concepts is at least roughly the same thing as what Rudolf Carnap speaks of as "explication" in the sense of "transforming a given more or less inexact concept into an exact one or, rather, replacing the first by the second" (Carnap 1950, p. 3). See also the rest of Carnap 1950, pp. 3–8.

description we are talking about it. (Besides chapter 2, levels of description are defined and described more thoroughly in Appendix A.)

Meanwhile, determinism is an obviously central concept. I make sure to give a definition of it that covers everything essential about the existing debate and especially the randomness argument that I give later. The essence of the definition is that determinism holds of a particular part of the universe on a particular level of description if natural law or other rules governing the world on that level leave only one possibility as to what will happen, given a certain starting condition. Indeterminism is just the denial of determinism. (The definition for determinism and its consequences are discussed in more detail in Appendix B.)

The rest of Chapter 2 acts as a background so that I can later refer to common arguments in the debate, even if they are not central to the argument of this work. I start by examining to what extent incompatibilism is intuitively appealing, noting that it is in some ways, but the picture is far from simple, and there is a tendency for people to speak as if there is some third option besides determinism and indeterminism (even though that should be definitionally impossible) that applies to free will. Then I look at different aspects of what has been considered as important for free will, and follow up by asking *why* free will is considered important anyway. After this, I briefly go through many arguments for and against the compatibility of freedom and determinism. Finally, I examine some ideas about the incompatibility of free will and determinism that I think can be dismissed as misunderstandings.

Chapter 3: The Randomness Argument Against Libertarianism finally dives deep into the debate about free will and determinism, presenting the *randomness argument* against a libertarian conception of free will. To summarise, libertarians believe that indeterminism is a necessary requirement for freedom. The randomness argument, variations of which have been used many times before, is here given as a conjunction of two closely related arguments. The first is that indeterminism inevitably contradicts an important, well-defined sense of freedom or control that I call *concrete control*, thus suggesting that indeterminism is *in*compatible with freedom. The second part of the argument is that, largely due to

the contradiction with concrete control, there is nothing about indeterminism that is desirable as a requirement for freedom other than in ways that can be reduced to saying that indeterminism simply *is* a requirement for freedom. The chapter looks at many different libertarian theories and arguments and concludes that none of these can avoid the randomness argument.

An important point brought up as part of the argument is that, though there are different variations of determinism and indeterminism, given the way I have defined them – which captures something essential about the existing debate – there is no option that is not either. Much of the rest of the chapter is then dedicated to showing how this means there is no escape from the randomness argument, and how many attempted ways of avoiding the argument have only obfuscated the core point about determinism by referring to something fundamentally irrelevant, such as variations of causality. After getting creative enough in the search for ways to subvert the randomness argument, I do find some ways in which randomness can be independently useful; however, they do not require the libertarian indeterminism on the ultimate level, and they do not otherwise match most of the libertarian motivations for wanting indeterminism.

Even the seemingly different libertarian requirement of being the ultimate origin of one's own choices is shown to amount to the same thing as the requirement for indeterminism. To be an ultimate origin of anything requires a self-creating act, but since there is no self prior to this act, it cannot be under the (concrete) control of the agent, and thus the best the believer in self-creation can do is to postulate that some random events are the acts of the agent (which I call *postulated ownership*). Since indeterminism is also postulated as a requirement for ultimate origination, indeterminism becomes the only real requirement, and conversely there is nothing to be gained from such ultimate origination other than indeterminism.

The randomness argument with all its facets is also summarised in Appendix C.

Chapter 4: From the Randomness Argument to Compatibilism goes on from where the previous chapter left off to use the randomness argument to argue for

compatibilism. It is shown that there are two logically possible but mutually incompatible views of freedom, representing our different intuitions about it. One is compatibilist – even requires determinism – and the other is libertarian. The libertarian version offers indeterminism, but little else. The compatibilist version offers, roughly speaking, everything except indeterminism; it is what makes concrete control possible. Since the two prongs of the randomness argument showed that indeterminism can be harmful to freedom but cannot be desirable for any other reason than the intuition or theoretical claim that it is necessary, it is easy to conclude that the compatibilist conception of freedom is much more desirable and should be adopted. If we consider incompatibilism a misstep and forget about it, everything else falls into place neatly; there are no unpleasant repercussions.

Chapter 5: Beyond Determinism: The Special Nature of Freedom goes back to the question of why, since there is no reason beyond intuitions to accept incompatibilism, it *seems* as though freedom and determinism contradict each other in the first place. This starting point is used to make a number of observations about what free will really means and what it requires besides determinism.

Firstly, two thought experiments illuminate how we should think about deterministic free will, and why this leads to surprising results we might not expect under determinism. The *decision machine* asks us to imagine a conscious, deterministic chooser in a deterministic universe. It points out that it is natural, perhaps even necessary, for the chooser to see different options open to it before it has made its decision – even though the scenario assumes that the choice is deterministic. The *prediction machine*, meanwhile, imagines a vast computer sitting outside our universe but connected to it and able to predict everything happening within the universe. It is argued that if the computer's predictions are communicated into the universe, they may by perfectly normal causality change what happens in the universe, thus making it a more complicated task than it appeared to state what it could mean that the predictions will always be true. Deterministic choices of agents within the universe are one means by which the predictions may come to be falsified, and further, if an agent is not able to effect such a falsification, then depending on

the details, this may signal that the agent is not acting freely. Thus, it is shown that at least *some* kind of rigid predictability is in conflict with free will, even when determinism itself is not.

Next, I present something between a psychological hypothesis and an observation about the philosophical discussion on free will, concerning the kind of thinking behind incompatibilist intuitions. It seems that there are three psychologically incompatible ways of thinking about causality, agency, freedom, responsibility, and so on. In the *mode of randomness*, things are seen as happening for no (meaningful) reason. In the *mode of causality*, things are seen as happening in a rigidly deterministic way that no-one has any choice about. In the *mode of agency*, actions are seen as happening due to agents' intentional choices (and a full causal analysis of them is suppressed). It seems as if these modes of thinking are the reason people oppose compatibilism but also gravitate towards a supposed third option rather than seeing the consequences of indeterminism. Thinking in the mode of causality is incompatible with thinking in the mode of agency, and this is mistaken for an incompatibility between determinism and freedom. In reality, though, my arguments up to this point have shown that the way the mode of agency is understood in practice is more deterministic than anything, and that indeterminism should be modelled under the mode of randomness. However, the three modes are apparently powerful psychological attractors. Also, it is true that the particular kind of determinism imagined in the mode of causality is incompatible with free will. This gives us important hints as to what free will is *aside* from being deterministic. I start to answer this question in the final major section of this chapter.

If *some* kind of determinism contradicts freedom, it is important to know which kind. I reason that it is determinism of such a kind that does not allow us to make a different choice even if we have reason to. Thus, it is not about being "able" to do otherwise for no reason, as in indeterminism; it is about being able to do otherwise *given* a good reason, which is as compatible with determinism in the broad sense as anything else. I argue that this means that free will is incompatible with determinism on any higher level of description that may leave out any reason the

agent might have to choose otherwise. If you can make any generalising remark about how the agent will behave in a certain kind of situation, that means the agent is not free to act differently in that sort of situation even if there is a reason. A way to describe the core idea of such compatibilist freedom turns out to be *reasons-responsiveness*, which could be summarised as "the propensity to do that which there is good reason to do."

In **chapter 6: Moving from Freedom to Responsibility – and Punishment**, I finally turn to look at the question of moral responsibility. I introduce it through the works of two other authors: Göran Duus-Otterström's *Punishment and Personal Responsibility*[23] and Saul Smilansky's *Freedom and Illusion*[24].

In *Punishment and Personal Responsibility*, Duus-Otterström gives a useful analysis and introduction of the concept of moral responsibility. In my summary of the book, I repeat some of this analysis to show what is typically assumed about responsibility. The book also serves as an introduction to common assumptions in that Duus-Otterström takes it for granted that what he calls "libertarian free will" is what is required for responsibility, arguing from intuitions that he only partly acknowledges to the conclusion that it is best to bet on the existence of such free will.

Smilansky's *Freedom and Illusion* shares the basic assumption that "libertarian free will" is required for true responsibility. However, Smilansky does not take other things for granted. He understands the basic problems with ultimate origination and randomness that I explain in my chapter 3, and thus, he concludes that since "libertarian free will" is impossible, so is proper responsibility. He acknowledges that there is some lesser value in a compatibilist concept of freedom and responsibility, though, so he concludes that it is just as well people are mostly under an illusion that they can have real freedom and responsibility, since he believes

[23]Duus-Otterström 2007.

[24]Smilansky 2000.

seeing the truth would keep most people from properly respecting what we do have.

After introducing these views, I question them. I take detour to argue about how intuition cannot be taken to prove many things and then go back to show how this is a problem for Duus-Otterström an Smilansky. In particular, I show that they both depend inexcusably heavily on somewhat veiled appeals to intuition; both Duus-Otterström and Smilansky take it for granted that "libertarian free will" is true freedom, that it is necessary for responsibility, that it is *sufficient* for responsibility (with some caveats, of course), and that responsibility justifies punishment. As for the connection between libertarian free will and responsibility, I argue that it is a dead end of absurdity; considering that (in the form used by both authors here) it describes something contradictory and impossible, how could it be that it is a vitally important moral category? This appeal to intuition on my part aside, there is simply no evidence for the assumption that "libertarian free will" is either necessary nor sufficient, aside from intuition, which is not really evidence here. Meanwhile, the notion that whatever responsibility means, it justifies punishment, is a very strong moral claim – the claim that it becomes morally right to hurt some people – that, I argue, would take more to prove it.

In **chapter 7: Deconstructing Responsibility**, I lay bare and make explicit common assumptions about moral responsibility – and then question them. I start by presenting the common assumptions as a scheme where different concepts are connected to each other, basic assumptions such as that responsibility implies free will and the appropriateness of praise, blame, and punishment, that freedom requires alternative possibilities, and that certain persons such as children are at least partly exempt from responsibility. An important thing that is revealed is that, how we may psychologically feel about it notwithstanding, responsibility largely means just that it is appropriate to treat the person who is responsible in certain ways – notably punishment, which I largely focus on in the rest of the chapter.

It is evident from both common experience and psychological studies that people tend to take it for granted that wrongdoers should be punished, just because they have done something wrong. As stated above, though, this is not something that

can be established as a moral truth by appeal to intuition. Considered objectively, this *intrinsic-good retributivism* appears as baseless and morally wrong: if we are normally obliged not to hurt people, how does that imperative suddenly get reversed?

I support this attack on retributivism through both social psychology and evolutionary psychological observations. Firstly, a commonly accepted explanation from evolutionary psychology explains the retributive instinct as a natural adaptation effectively derived during evolution from its consequentialist benefits. Though this does not prove that the retributivist intuition is not morally correct, it raises further warning signs about the idea that the intuition should be treated as anything more than a psychological fact. Secondly, there are results in social psychology that show the delusion and harm often (though not conceptually) involved in retributivist attitudes. Though this does not prove that the philosopher's perhaps cool retributive intuitions are the same, it raises still more warnings.

This work does not advocate abandoning responsibility or even punishment, however. **Chapter 8: Reconstructing Responsibility** establishes why the concept of moral responsibility, with the very schematic features that it was described as having in the previous chapter, is in fact an important moral concept. In fact, I argue, my way of deriving responsibility from more fundamental moral values and principles of rationality presents it as a more important part of our moral lives than if it were a free-floating absolute value based on intuitions.

I start this chapter by discussing kinds of reasons-responsiveness and introducing my own definition that is somewhat vague but works better for practical reasons just because of that. After this, I present the key piece to the puzzle of why responsibility is associated with free will. If the implicit purpose of being held accountable – praise, blame, punishment – is to encourage and discourage behaviours, then there is no reason to hold anyone accountable for things such that they could not choose to do otherwise for the reason that they know they will be held accountable. In other words, there is no point in holding people accountable if they are not suitably reasons-responsive – and, as such, suitably free. There is good reason, then, to consider it unjust to hold someone responsible for something they

could not help. If someone is "pathological" enough that they would not be deterred by punishment, it makes sense to give them treatment instead, effectively trying to restore their reasons-responsiveness. The ways we actually and historically have judged people to be responsible or not fits this logic very well, even though the reasons are often articulated in different terms.

There are some questions left to answer before the final conclusions. It is not enough to say that the free person is reasons-responsive, it must also be explained *which* reasons are the good ones the person should be responsive to. I discuss an example of views about the "real self" and argue that this approach is insufficient. Instead, I argue (following Mary Midgley's[25] explanation about the nature of morality) that, in judging which values and motives are good ones, we must consider them and their environment as a whole. Aside form basic pleasure (or pain), only those things can be good (or bad) that promote harmony (or the opposite) within the person as a whole and between them and other persons and impersonal natural facts.

In **chapter 9: A Unified Theory of Freedom and Responsibility**, I bring it all together and finally give the definitions of free will and responsibility that I have been working to solve, along with with various thoughts and considerations about what follows from them. I start off with getting some perspective to how to understand the nature of the free will question, and how to understand that we can be free while being influenced by the rest of the universe. I try to paint a picture of how it is fine to be part of the universe rather than solipsistic chooser.

Free will of the sort that is worth wanting is defined as an ability to rationally, reasons-responsively make choices based on good reasons all things considered. The *challenge of freedom* is that we can never be perfect in this respect, indeed we may not even be very good at it to start with, and we need to strive to become more free. Nevertheless, a certain "normal" amount of freedom is sufficient that it is sensible to hold people morally responsible.

[25]Midgley 1994.

Meanwhile, responsibility (in general and moral responsibility in particular) can be reduced to the opportunity to bring about good consequences. If someone can reasonably bring about good consequences by doing something, or avoid bad consequences by not doing something, there is a moral reason to ask or demand them to make the choice that will have the better consequences. This also gives reason to give the person feedback afterwards, also serving as feedback to others in a similar situation, encouraging better choices overall. Since responsibility demands opportunity, a lack of freedom – of will or otherwise – about the choice removes responsibility just as it removes opportunity. The *challenge of responsibility* is that, in applying the concept of responsibility in ever new kinds of situations, we need to know when the concept is appropriate to apply – when it encourages choosing the better option – and when it is unhelpful and only serves to cause useless anxiety, guilt and so on.

With these answers, we can see that the assumptions about freedom and responsibility, and the schematics they form, can be "filled out" with the details of the theory I present here. For example, if responsibility requires freedom of will, and freedom of will requires alternative possibilities, we should be able to continue thinking in these terms while understanding "freedom of will" and "alternative possibilities" in terms of reasons-responsiveness.

Lastly, I look very briefly at what science says about our degree of freedom by the definition that I give, and consider what we should think about the results. I come to the conclusion that our freedom is woefully limited, largely due to irrationality and the effect of unconscious impulses, but it is also possible for each of us to develop it further. The much-cited Libet experiment[26] may be nothing more than a red herring – all it perhaps shows is that we may make decisions before we are consciously aware of it, but there is no need to identify ourselves with only our conscious sides. Nevertheless, unconscious motives are potentially threatening,

[26]Libet 1985.

because if we do not act on good reasons and are not even aware of what our real reasons are, we are threatened with ignorance, delusion and a lack of reasons-responsiveness. Psychological research shows that there are plenty of ways in which this happens.

In the end, I conclude, we find ourselves in the position of imperfectly free beings with the responsibility to become better, with no certain answers but plenty of opportunities.

2 A Brief Introduction to the Debate about Free Will and Determinism

> "[When] people are talking about free will and they start talking about determinism versus indeterminism, you instantly know not to take them seriously. This is not what the issue is."
> -Sean Carroll[27]

The debate about whether determinism is compatible with free will or not, between compatibilism and incompatibilism, has been going on for a long time and taken innumerable turns, even if these have mostly led to going around in circles. My main contribution to the compatibility debate in chapters 3 and 4 will seek to straighten out many of these curves by pointing out some simple basic arguments that are unaffected by the more complex arguments. Before that, to map out the terrain, this chapter takes a hurried tour of the existing arguments for and against compatibility, as well focusing slightly more on some observations about the debate that will become relevant later.

First, I take a look at how the concept of determinism is defined in this work, as well as the concept of *levels of description* that I use together with it. These are discussed in more detail in Appendices A and B, but I give a brief introduction here to make sure the reader has the general idea.

Next, I look at what our intuitions say about the incompatibility or compatibility of free will and determinism, which turns out to be anything but

[27]Carroll 2021.

simple. The next part looks at why it is that free will is thought to be important and desirable in the first place. After that, I give a very brief overview of numerous arguments for and against (in)compatibilism. Finally, the last section presents and refutes arguments for incompatibilism that are, I argue, simply mistakes about what determinism or other concepts mean, and thus can be defused easily without needing to appeal to the major argument presented in chapter 3.

2.1 Determinism and levels

What actually is determinism? As one might expect, there is no one definition uniformly used by everyone, and the questions of definition can get quite complex. This also causes some extra difficulty in discussing determinism and free will, as one must be careful to know which definition they are using, and different parties may use different ones. Fortunately, the discussion in this work does not need to concern itself with multiple different definitions. Quite the contrary, the arguments about determinism, indeterminism and free will that are seen as central here are best addressed by sticking to a relatively simple core idea of determinism.

In this section, I introduce the idea of determinism that I will be using, as well as an accompanying idea of *levels of description* that will be needed to make questions of determinism and agency more specific. In the interests of getting to the topic of free will itself sooner, I will keep this shorter than it might have been, but in the interests of making sure the definitions are clear and do not give rise to misunderstandings and counterarguments based on equivocation, I elaborate the definition of levels in Appendix A and the definition of determinism in Appendix B. I see considerable confusion in the existing debate with respect to what is being meant by determinism and indeterminism and what follows from them, so I recommend paying heed to those appendices as well – especially when formulating counterarguments, so as not to go around in already trod circles. In this section, I will keep the definitions brief and focus more on getting across an impression of

what I mean.[28]

The definition of determinism I use plays an important key role in my argument, but at the same time, it does not aim to answer all possible questions about determinism. The idea is that the definition should be detailed enough that there will be no holes in it leading to unanswered questions but, at the same time, broad enough that it will not fail to apply to any relevant cases just because of the wording.

Not everyone whose ideas I discuss has used the same definition. The point, as I will argue, is that certain arguments work with this definition, and those arguments will continue to apply to different possible cases and theories regardless of what vocabulary is used in stating those cases or theories. So, for example, if someone ties their definition of "determinism" to causality in a way I do not and seeks to solve problems involving "determinism" in a way that involves invoking the notion of causality (see e.g. section 3.4.2), I can still ask whether their solution solves the problems I have identified using my definition of "determinism".[29, 30]

[28]Chapter 3 also illuminates the concept of determinism that I use, because it applies it to different aspects of the debate repeatedly and in different ways.

[29]Cf. Earman 1986, pp. 5–6, which argues against using the definition of *determinism* that it means that every event has a cause. See also Slattery 2014, pp. 118–121. Cf. van Inwagen 2002, pp. 65.

[30]There is a possibility, at least hypothetical, that since I define *determinism* in this particular way that is not shared by everyone, I may end up e.g. classifying someone as a "compatibilist" or "incompatibilist" on the basis that they (do not) agree that "determinism" as they define it is compatible with freedom, but they would not agree to that statement about *determinism* as I define it. I try to note where others use words differently than I do, and to follow the consequences in my own terminology, but it is possible I might miss such a difference. In particular, I address the question of "libertarians" that are not really incompatibilists in my sense in 3.2.3.
That said, this potential problem is hardly specific to my approach of using a specific definition; if anything, using definitions loosely would be more likely to lead to problems of equivocation.

2.1.1 The Laplacean demon

A classic and frequently used[31] image of determinism goes back to Pierre-Simon Laplace, particularly this quote:

> We ought to regard the present state of the universe as the effect of its antecedent state and as the cause of the state that is to follow. An intelligence knowing all the forces acting in nature at a given instant, as well as the momentary positions of all things in the universe, would be able to comprehend in one single formula the motions of the largest bodies as well as the lightest atoms in the world, provided that its intellect were sufficiently powerful to subject all data to analysis; to it nothing would be uncertain, the future as well as the past would be present to its eyes.[32]

Such an intelligence can be called the *Laplacean demon*.[33] This gives an idea of what kind of thing we are talking about, but I will not build my definition on quite this basis because I do not wish for "determinism" in the sense intended to be regarded as an "epistemic" term. I aim to give a description of how the world might be, not what could be "known" about it – at least insofar as that is in any way a different question.[34, 35]

---

[31]Robert Sapolsky states sarcastically that it is "virtually required" to invoke Laplace when starting to explain determinism (Sapolsky 2023, p. 15). I merely find it convenient.

[32]Cited here from Nagel 1971, p. 281, where it is apparently Ernst Nagel's own translation from Laplace's *Théorie analytique des probabilités*.

[33]I take the term from Chalmers 2012, p. xiii.

[34]I would not use the word "know" in a definition in any case, since I believe it is itself a fundamentally undefinable term of human convenience, implying that any definition of a concept tied to it would not give clear criteria for when the concept applies. *Undefinable* here does not imply that it cannot be explicated, but that the way it is used does not imply a proper definition, and thus the correct explication does not take the form of a definition but something else. I am working on an article on this claim.

[35]John Earman (1986, pp. 6–8) dismisses Laplacean predictability as a definition of "determinism" on the three grounds that introducing a knower complicates things in terms of what effects the knower's abilities may have on things, that epistemology

2.1.2 Levels of description

One important tool we will need in discussing determinism and free will is the idea of several "layers" of reality somehow "built atop" one another. We will need to ask questions such as what is the relationship between determinism or indeterminism on the fundamental physical level and choices on the human level on which we perceive things, and talk about something being deterministic under one level of detail of description but not another. Though I do not go into much detail about such things, some definitions will be needed so that such talk is not uselessly loose, and some explanations so that it is comprehensible.[36]

I take an extended quote from Christian List to explain the general idea of such levels, simply because he articulates the point so well:

> Facts about the world—in general, not just in relation to human agency and free will—are stratified into levels. In the sciences, "levels" correspond to different ways we describe the world. To make sense of the basic laws of nature, for example, we employ fundamental physical descriptions, drawing on our best theories of physics. We use concepts such as particles, fields, and forces, and specify a variety of equations that describe how physical systems evolve over time. By contrast, to make sense of chemical or biological phenomena, we need to go beyond fundamental physics. Molecules, cells, and organisms all display patterns and regularities that can only be captured at another level of description, using a conceptual repertoire distinct from that of fundamental physics. As philosophers of science have pointed out, even a property as simple as acidity cannot be satisfactorily described in fundamental physical terms alone. In the case of acidity,

---

and ontology are best not confused, and that since prediction already has a name, calling it also a type of "determinism" does not make things any clearer. See also *ibid.,* p. 64.

For another problem for determinism as predictability, see 5.2.2 in the present work

One discussion that does equate determinism with a kind of predictability is found in Honderich 2002, pp. 81–83.

[36]This topic is discussed in more detail in my Master's thesis (Kokko 2014).

there is no easily describable configuration of fundamental physical properties that exactly matches this chemical property. In particular, there is no "translation scheme" that fully translates talk of acidity into talk of particles, fields, and forces alone. We need the concepts and categories of chemistry to talk about acidity. And we are here still dealing with a fairly basic example, compared to other, more complex chemical or biological phenomena.

In case you are in doubt about the need to go beyond fundamental physics to make sense of the world, try to explain cell division, genetic inheritance, or evolution by referring to nothing but molecules, atoms, and other elementary particles. Each living cell consists of billions or trillions of atoms, and an organism consists of billions or trillions of cells. Even the best supercomputer would struggle with the astronomical task of computing the processes inside a single organism at the atomic or molecular level. And even if, against all odds, these difficulties could be overcome, perhaps with the help of massively distributed computing on the internet, then a purely microphysical description of cell division, genetic inheritance, or evolution would still fail to pick up many macroscopic regularities we are interested in. Indeed, such a description would not help us to understand the relevant phenomena at all. It would make us lose sight of the forest for the trees. Now, once we turn to the domain of humans and their intentional actions, fundamental physical descriptions are wholly inadequate. In the language of fundamental physics, we cannot even talk about tables, trees, and other ordinary objects—only about particles, fields, forces, and so on. If we wish to make sense of people and what they do, we require psychological descriptions—descriptions that refer to thoughts and beliefs, preferences and desires, goals and intentions.[37, 38]

[37]List 2019, pp. 5–6. Numbers referring to endnotes removed.
In this book, List gives a great deal of attention to levels of description in relation to determinism and free will. I applaud his understanding and explanation of levels of description, but I disagree with crucial points related to the implications for free will; see section 3.4.3 in the present work. Note that List was also mentioned in section 1.3 as the proponent of "libertarian compatibilism".

[38]For another explanation of why different levels of description are needed, see Williamson 2020, pp. 11–12.

I see these levels as being true levels of reality as well as of description, and thus ontological, although in a weak sense, since their relationships are still based on weak emergence. In weak emergence, "higher" levels have some independence of the "lower" ones that they are built on, but the higher cannot exist without the lower, nor do the new properties manifested by the higher level include anything that cannot be *in principle* be explained from the lower level. If it were to turn out that the levels are connected by strong emergence instead, it would only make it easier to speak of how the levels are effectively separate, as strong emergence allows for separation that does even need to be explicable.[39] Conversely, if you do not accept the different levels as I describe them as having the status of being ontologically separate, this does not affect my argument. The way they are described here suffices for saying that they have the consequences I posit.

2.1.3 Determinism and indeterminism on different levels

To speak of determinism or indeterminism is always to speak of it on some level of description. At the same time, we could say that the universe just is (in)deterministic, as it were at the bottom. As I understand it here, this is the same as to say that it is (in)deterministic on the *ultimate level*, a presumed "theory of everything" level on which every other level is built.[40] If everything else is weakly emergent on something else, and there is no infinite regression or circularity of levels, this level is the only one that is not emergent. I assume that there is such a level in my discussion, but if there is not, that only means that we have to make do with speaking of non-ultimate levels, and this discussion will be limited to what I say of them.

It may be surprising, but weak determinism allows for the possibility of higher levels of description being indeterministic even though their lower levels or the ultimate level is indeterministic, as well as for the higher level to be deterministic

[39]For more on the sorts of emergence, see 10.2.

[40]See 10.6.

when the lower level is indeterministic. Very shortly put, this is because information from the lower level is lost on the higher level, and thus one state on the higher level can correspond to more than one state on the lower level (a phenomenon known as *multiple realisability*). This is discussed in detail in section 11.3 in the appendix.

When speaking of different positions with respect to free will and determinism, such as compatibilism and incompatibilism, the default will be to assume that the (in)compatibility is with determinism on the ultimate level. However, since the actual choices are inevitably conceptualised on a higher level, discussing other levels and especially the psychological/intentional level of choices will be very relevant as well.

2.1.4 This work's formal definition of determinism

As detailed in Appendix A, the full definition of *determinism* used in this work is the following:

> Determinism holds in a part of the universe P at a level of description L between times T_1 and T_2 iff, given any total state of part of the universe P as described in the terms available in L at T_1, and any total input from the rest of the universe between T_1 and T_2, as described in the terms available in L, only one total state of P as described in the terms available in L is possible by the rules of L at T_2.

I will unpack this here very quickly to bring out the points embedded in it, though obviously the appendix explains this in much more detail.

Part of the universe P*:* Part P may be the whole universe or a proper part[41] of the universe, so this covers both determinism applying to the universe as a whole or

[41] In philosophical parlance, the whole thing can technically be considered as a part of the thing (such as the universe being a part of itself), whereas a proper part is a part that is not the whole thing of which it is a part.

only some part of it. “Part” may be just about anything, as there seems to be no particular problem resulting from not restricting its definition. Thus, some things that count as parts of the universe might be areas of space, particular individual events, or particular continuing processes.

Input from the rest of the universe. If we are only talking about a proper part of the universe, then we have to account for the fact that the rest of the universe may have a causal effect on it. Hence, it might be that the initial conditions of *P* do not entirely determine what happens in *P*, simply because something from the outside interferes. The way I am interpreting determinism for a part of the universe here is that interference from outside *P* is effectively considered the same way as the initial conditions of *P*, even though it may come later: the laws governing *P* can still be deterministic as long as they react to that input in such a way that given the same input and the same initial conditions, the result will always be the same.

On level of description L*, in terms available in* L. We are always talking about determinism on some level of description or other. We are always doing this even if we do not consider it. However, if we speak of the universe being deterministic, we usually mean it is that on the ultimate level.

The rules of L. This is what I also call *laws of nature**: the rules that govern the world or part of it according to the description on the level *L*. These may count as actual laws of nature or not, as not every level of description includes rules that are seen as laws of nature. An obvious example of such rules in the context of free will would be a psychological explanation, invoking general concepts, of why a person does what they do and what they are likely to do.

2.1.5 Why is there no need to ask whether determinism is true?

Considering all the talk about determinism and its consequences in this work, it may seem strange that one thing I do not comment on is whether the universe actually is deterministic or not – or even whether some parts of it are. There are two main reasons for this; I focus on the first one here and introduce the second one better in

section 9.1.3 with the benefit of textual hindsight from having done the rest of this study.

The first reason why I am not asking whether determinism is true is that I am looking at what it would mean for free will and responsibility *if* it were (or were not, or it were partly true). A big part of the debate has always been what the relationship actually is between freedom and determinism – and it should have been an even bigger part of the debate considering that sometimes the answer has been taken for granted when it should have been questioned. Answering this question is in principle necessary for even knowing what to think about it if determinism *is* true or is not. We would not know the implications for determinism otherwise. A large part of the goal of this work is to figure out what determinism *would* mean for free will, not whether it is true.

Once we know what to think of free will and what it means, we can start asking questions about whether we can or do have it. As it will turn out, there are multiple reasons why the question of determinism (especially on the ultimate level) is not the most relevant thing to ask about at that point. These reasons are what will be summarised in section 9.1.3.

It would still be relevant to ask how much free will we have in terms of the notion of free will I will propose, not in terms of the physics of determinism and indeterminism, but through something like the lens of the empirical psychology of decision-making. To do that in much detail goes beyond the possible scope of this work, but the question is briefly touched on in section 9.6.1.

2.1.6 On intentionality, materialism and dualism: An important note especially to libertarians

Since determinism is often seen as physical determinism, the question of free will and determinism may be associated with questions about whether an entirely materialistic, physicalistic or mechanistic universe can contain those human qualities that are needed for freedom. Can intentionality be derived from a mechanical universe? Does real personhood require dualism in the form of the soul or

consciousness or some equivalent of these being something non-physical?

Though these themes will be touched on occasionally (see 3.4.8), this dissertation is not addressing this point. I admit my own view is that just about everything can be explained in terms of weak emergence from a physical basis. I think that there is nothing to explain about intentionality that cannot be explained this way, and I am optimistic even about solving the hard problem of consciousness without dualism. However, none of this is required for my argument.

I have noticed that when incompatibilists and libertarians hear talk about determinism in the context of free will, they can easily jump into imagining more than is perhaps being said: that determinism implies a physicalism (etc.) that they find to be incompatible with their notion of free will. While determinism is often associated with things like physicalism, it does not have to be. In this section, I have introduced my own definition of determinism that I will be using in the present argument – and I want you to notice that it has no mention of things like physicalism, mechanism or monism. As it is still determinism, I know it may sound mechanistic, but I ask that you give it a chance. The point is that my argument is so general it is independent of metaphysical theses like physicalism.

If you think consciousness needs to be added to the physical world in a dualistic way, I can grant that may be the case. If you think intentionality is fundamentally different from causality or physicality and needs to exist separately from it, I can grant that may be so. If there is anything you want to add on top of the physical world, be my guest. You can read all of the following assuming the physical world is not enough to cover what there is and these other things exist. I just ask that you reflect on the definitions of determinism and indeterminism that I introduced (and elaborate in the appendix), as applied to those other realms as well as the physical, and then ask the questions I ask and consider my arguments. In other words and for example, if you think minds are immaterial, I want you to walk along with me and ask questions about what follows if immaterial minds are indeterministic or if they are deterministic – in the basic sense that is independent of materiality and immateriality.

If you can do this instead of jumping to imagining determinism as incompatible with free will because of physicalism, then we can all be following the same arguments. If the threat, to you, is physicalism, you can rest assured none of these arguments are supposed to defend physicalism or depend on it. Next, you can judge for yourself whether they are, for example, too mechanistic – just hear them out first.

A further illustration of this theme is found in the appendix B (in 11.9). There, I show how Thomas Reid speaks in terms that concern things libertarians are concerned with, and opposes a form of determinism, but still describes free will in compatibilist and determinist terms. For libertarians who sincerely want to understand what I am saying and give it a chance, I highly recommend paying attention to that subsection as well as the present one.[42]

2.2 Intuitions and incompatibilism

I think that the idea that incompatibilism is based on common intuitions about free will is hard or impossible to contest. However, that does not mean that it is the only view derivable from common intuitions, or the one that should be adopted if we use intuitions as the basis of our view.[43] Intuitions are not coherent theories, and as such, building theories based on them is a matter of interpretation, as well as adaptation in somewhat the same sense as adapting a book into a movie. When looked at fairly and more closely, intuitions are revealed to be something much more messy than theories.

There are two ways of examining "common intuitions" about free will and

[42]On the general topic of a dualistic/spiritual/religious etc. vs. materialistic/scientific etc. world view and the question of free will, see also Balaguer 2014, pp. 3–7 and chapter 3. On the other hand, that whole book is a good example of treating the issue in a way that takes into account both physicalistic and non-physicalistic possibilities and applies variations of the same arguments on both.

[43]This is not to be taken as implying that we *should* base our views on the best definition for free will on intuitions. See especially 4.3.3 (p.113) and 6.3.1 (p.159) below.

determinism. The first is a more limited and "philosophical" approach, where a philosopher can study the philosophical literature and their own intuitions[44, 45]. The second is empirical study of laymen's intuitions.[46] From my point of view, as will be seen below, results of the two approaches handily converge: both reveal roughly the same contradictions in intuitions.

That said, there is obviously no universal set of intuitions. I am merely making the claim that intuitions such as I describe below – with references to my own intuitions, the statements of other philosophers, and empirical studies – are somewhat common and possibly even dominant among Western people (as discussed in the next section). All it takes for my thesis later (mainly chapter 5) is for them to be, shall we say, robustly existent – existent and not completely marginal. Even if they should turn out to be marginal after all, which my evidence is enough to make unlikely, that would be more of an embarrassment than a real refutation of my points.

2.2.1 Contradicting intuitions

It is easy enough to evoke incompatibilist feelings in many people. That group of people certainly includes myself, even though I do not hold an incompatibilist theoretical position. If you tell a person to whom this applies that determinism means that when they make a choice, the universe is in advance such that only one choice is possible, they will (usually?) feel like that means there is no free choice. You can get the same kind of result by taking the longer perspective and saying that everything they do was already determined by the initial conditions of the universe, or generally factors before their birth.

[44]Cf. Nichols 2006, pp. 61–63.

[45]Obviously studying one's own intuitions and studying the philosophical literature could be called two different approaches. They do have something relevantly in common, however, in opposition to what I define as the other approach here, so this coarse-grained (see 10.5 for that term) grouping works fine here.

[46]See Nichols 2006, pp. 64–65.

Such intuitions translate more or less directly into incompatibilism. Though it is still a matter of interpretation, it is definitely the obvious, most straightforward interpretation. Conversely, a compatibilist will need to dismiss such intuitions as false – or perhaps find a more convoluted but still sensible interpretation for them, which is something I attempt in chapter 5.

Thus, we can say that there are common intuitions that lead straightforwardly to incompatibilism.[47, 48] This is where things get complicated, however, because there are also common intuitions in the opposite direction. The simplest such intuition is a general assumption of something like determinism. Thinking about how things work in the world naturally leads to thinking they happen for a reason, meaning they are determined.[49] When this is not taken as a position of hard incompatibilism, it suggests compatibilism: if both freedom and determinism are believed to be actual, that implies they should be believed to be *possibly* both actual at the same time too, that is, compatible.

Philosophers definitely have different intuitions. What about what studies reveal about the intuitions of people in general, potentially untainted by theory? This is a topic that could be studied in detail in the context of the present discussion, but it would be too much in terms of the scope of this work. Thus, I will summarise what the research has shown in a coarse-grained way that says just enough for the purposes of my thesis later: Many studies have been conducted, finding different results on

[47]Cf. also Earman 1986, p. 239, Honderich 2003, p. 93–95, and Vargas 2013, chapter 1, though this last source could equally be used to support the point that intuitions are contradictory.

[48]Contrast this with Alan Donagan's (1987, pp. 179–182) contention that we learn to presuppose we could have done otherwise (in what he takes to be a libertarian sense), and this is commonly believed, but it cannot be derived from our experience without this belief. Whether this should actually be considered a different view or a variation of the same view about common intuitions is not obvious.

[49]See e.g. Reid 2010, pp. 26–27; Honderich 2002, chapter. 2 and p. 145; Balaguer 2014, pp. 12–14; Honderich 2011, p. 444.

whether people seem to be naturally compatibilists or incompatibilists, depending on how they are asked. Different theories have been proposed and tested to explain what is really involved in common intuitions, but no grand overarching explanation has been reached.[50] Thus, probably the most justifiable position is that intuitions in general are indeed contradictory and contain elements of both compatibility and incompatibility, both determinism and indeterminism as involved in free will. This point should be kept in mind especially if someone claims that intuitions support either incompatibilism or compatibilism; both claims seem to be partly true, but neither in the sense that that gives either position an advantage.

2.2.2 The appeal of a third option

An oddity about the debate between compatibilism and incompatibilism is that so many participants endorse or seem to implicitly assume, with varying degrees of implicitness and explicitness, options that are (supposedly) neither determinism not indeterminism. Since *indeterminism* has been defined as "the absence of determinism",[51] it seems as though there is no such third option. Whether we can speak of such a thing is thoroughly discussed in chapter 3, most directly in 3.2, but for now, it suffices to notice this tendency exists.

Other philosophers could, of course, simply have different definitions for "determinism" and "indeterminism" than I do here, and when they explicitly deny both, they could be talking about their definitions. To varying degrees, this is so. However, when they affirm a third option, they are not admitting that there are problems related to determinism and indeterminism by the definitions I use here, yet these problems would be problems in their theories as well. Thus, it is not just that the philosophers taking the kind of stance I am talking about here are talking about something else; they are also skipping the real issue with free will, determinism and

[50]Nadelhoffer et al. 2023.

[51]See 2.1 and Appendix B.

indeterminism that touches everyone – or they are acting as if they are addressing it when they are not. It is a reasonable guess that the tendency of philosophers to affirm or think in terms of a third option is somehow related to the tendency of common intuitions to be mixed between compatibilism and incompatibilism introduced in the previous subsection.[52, 53]

---

[52]Several of the philosophers discussed in this work show tendencies in this direction. Roderick Chisholm's paper on agent causality (2002) discussed in 3.4.2 contains the statement that the theory introduced there is neither "determinism" nor "indeterminism". There is also the idea, such as expressed by Laura Ekstrom (2011) as discussed in 3.4.1, that to equate indeterminism in choices with some kind of randomness is merely a mistake based on equivocation; this implies that there is supposed to be a third thing besides determinism and indeterminism of the sort that has an implication of randomness. The ideas of agents being akin to players in a game with indeterministic rules – see 3.4.8 – is also imagining a third option.

It is not always clear that one can fault a philosopher for making statements in these lines; Robert Kane may state that he aims to show that indeterminism does not need to mean randomness (Kane 2017, p. 2479; see the beginning of chapter 3 in the present work), but he is aware of the problems with this and makes much effort to get past them – though one might certainly say his efforts to do so reflect his missing the point that indeterminism just has certain consequences (see 3.4.7 and 4.3.1).

Some other examples: John R. Searle states that randomness is not freedom, but also dismisses compatibilism on the basis that what he means by free will is defined as incompatible with determinism (2007, pp. 44–45). He further writes that even though indeterminism in the case of the quantum events is "randomness", the physical system making up the person as a whole with its holistic properties could instead have non-random indeterminism weakly emergent from the quantum randomness (*op.cit*, pp. 75–76). Thomas Pink writes about how, under freedom as usually conceived, it is the case that in choosing between two options, both are possible, the agent determines which is chosen, and this is not random (2011, pp. 364–365).

Hard incompatibilists such as 'Trick Slattery (2014) and Saul Smilansky (2000) take it as a starting point that free will – often, the only real kind of free will – requires both determinism and indeterminism in an impossible way. Insofar as this is based on their own view of what free will really means, it means they also see free will as being a kind of third way, except that they realise that this is impossible.

[53]There is a spectrum rather than a clear line between tending towards assuming a third option and theorising that there are two (or three) different perspectives from which we view things as determined or intentional. For authors with the latter kind of idea, see 5.3.4. Even though I critical of wanting a third option, I will endorse the idea of such different perspectives later (see chapter 5 as a whole). My point in taking this critical stance here is to lead up to the idea that the idea of perspectives is a working way to understand what the intuition of a third option is perhaps "getting at".

2.2.3 The intuitive connection to responsibility

I will not really discuss moral responsibility until I get to chapter 6. At this point, I will simply note that there seems to be an automatic assumption – another intuition – that the same things that threaten responsibility threaten freedom.

It would be hard to discuss questions of freedom of the will without mentioning responsibility at all, at least when referring to what various individual thinkers have thought about free will. Incompatibilists about free will tend to assume responsibility is likewise incompatible[54], whereas compatibilists may make the contrary assumption[55]. I will make the occasional reference to responsibility as I talk about freedom and (in)determinism, but I will (mostly) refrain from drawing conclusions of my own about responsibility before switching from free will to responsibility in chapter 6. One major reason for this is that I do not intend to take the connection between freedom and responsibility for granted at all.

A little clarification to the above is in order. Among the meanings of "responsibility" are ideas about being the one who is responsible being the cause or origin of the thing they are responsible for. (See 7.1.) Notions of free will are connected with notions of agency, which is connected with being the origin of actions, and thus, there is a non-arbitrary conceptual connection between them and responsibility in *this* sense. However, moral responsibility is also seen as involving liability to things like praise, blame, rewards and punishments, and without including those, we are not really talking about *moral* responsibility (see 7.3). It is this latter connection that I refuse to take as a given.

Therefore, while we are talking about common intuitions, I simply wish to note the fact that there is commonly held to be a connection between free will and responsibility. At the same time, I will note that there are philosophers whose views

[54]For more discussion as well as references for this claim, see section 7.2.

[55]E.g. Dennett 2015.

do separate the two[56].

2.3 What is so important about freedom?

This work is going to be discussing different conceptions of free will and evaluating them against each other, also asking the question of which conception or conceptions we should adopt. To do this, we are going to need something on which to base our evaluation of the conceptions. There is also the general question of why we are talking about such questions of free will in the first place. For both of these reasons, this section looks at different reasons that have be given and could be given for valuing free will. This way, we can later compare the different possible conceptions of free will to these desired results of having it and see which conceptions can grant which of them.[57, 58]

This discussion will be fairly brief and cursory, since it relies on (presumably) common unanalysed intuitions about the importance of various things and their relationship with free will, and (much as with moral responsibility specifically, see 2.2.3) I do not wish to analyse these things as if they are known when they are only assumptions.[59]

---

[56]See particularly Fischer & Ravizza 2000.

[57]The main arguments referring to these desired features of free will and how different concepts of free will match them are found in chapters 3 and 4, summarised in section 4.3.2.

[58]Mark Balaguer (2014, pp. 7–9) contends that, when explaining all the reasons why we need free will, we might just be making excuses, with the real issue being that we *want* free will and it feels good to (think we) have it. While there might be something to this, I do not take it to invalidate the whole topic of discourse of free will being valuable for further reasons.

[59]For all the academic sources I have read, the summary in the podcast "The Free Will Show" (Cyr & Flummer 2020) was so superior in its presentation of a comprehensive list of reasons that the structure of this section is largely based on that. For more similar perspectives on the importance or supposed importance of free will, from authors with different views, see e.g. (libertarian:) Kane 2013, pp. 259–261, van Inwagen 1983, chapter V, (compatibilists:) Dennett 2015, chapters. 1 and 7, List 2019b, pp. 17–20, (free will sceptics:) Harris 2013, p. 5, Slattery 2014, chapter. 42.

Importance to, and of, responsibility. A primary argument for the importance of free will is that moral responsibility depends on it. If you are not free to choose what you do, perhaps expressed by saying that you are not free to do otherwise, then (the natural leap goes) you cannot be responsible for what you do. Our social and legal systems depend in the concept of moral responsibility, as arguable does the very way we relate to each other[60]. To abandon free will would seemingly lead to the need for a radical revision of our conceptions of responsibility and ways of understanding human agency – arguably to the point that many important systems would collapse and there would be nothing to replace them.

Admittedly, some (typically hard incompatibilist) authors instead argue that the notion of moral responsibility is a harmful misunderstanding and *should* be abolished, and this would lead to something better than the current system. This will be discussed later.[61]

The topic of appealing to moral responsibility as a ground for the desirability of free will is surprisingly problematic, particularly in the current context. In this dissertation, responsibility is discussed in chapters 6, 7, 8, and 9. This discussion will also answer the questions of how free will relates to responsibility and what kind of free will is helpful for responsibility. This also means that, unlike most of the other reasons for valuing free will that are discussed largely in chapters 3 and 4, the topic of what kind of free will enables responsibility is largely delayed until chapter 6 and later. Thus, the question of evaluating kinds of free will in light of their relationship with responsibility is also delayed until then. It will turn out, however, that the kind

Aside perhaps from van Inwagen, all of these authors purport to describe generally shared ideas about why free will is important or thought to be so, and they might succeed, but the possibility of their being biased by their respective stances on the question needs to be taken into account; I cannot comment on it in this space.

[60]Strawson, n.d.

[61]See 4.3.5.

of conception of free will I will defend as the best is also harmonious with the kind of responsibility I advocate as worth valuing.

Autonomy, authenticity, self-realisation, creativity. Free will is seen as a prerequisite for autonomy, "self-rule", being able to act of one's own direction and origination rather than, for example and in particular, being controlled by someone or something else. This is something often considered valuable, and perhaps it is. Free will is seen as helping with it or being a part of it, and perhaps that is so, too.

Another thing that people tend to value and is seen as dependent on free will is creativity. Perhaps only a genuine person with with autonomy and free will can be attributed as the creator of, say, a piece of art.

When we view ourselves from a perspective in which things we do or choose are due to something else than ourselves, it makes sense that this sense of autonomy could be threatened. Perhaps the world itself could just be like this, too, under any relevant level of description, rather than it being about the choice of perspective. It will turn out, though, that what kind of free will (if any) allows for autonomy is complicated and depends on the perspective taken – making the question what kind of perspective *should* be taken.

Agency and rationality. There's good reason for us to want to be rational agents, and free will can be seen as being required for this as well. For this, we need to understand reasons for acting one way or another, and also act based on them. Certainly some things seen as the opposite of free will would also be the opposite of this – such as being "programmed" to act only in a certain way. Other animals that behave more instinctually might not even count as agents.

The question of just what is required of free will to achieve the goal of rationality will be a very important one in my arguments. To discuss it here would be to get ahead of things; it is the topic of much of the discussion in chapters 3, 4, 5,

8, and 9.[62]

Virtue, temptation, struggle. We certainly tend to think that it is valuable when people do the right thing rather than the wrong thing, or for them to be virtuous of character in general. However, these ideas seem to imply a choice or an alternative. If you could not do otherwise, or were not making a choice to do otherwise, then maybe there would be nothing to praise you for when you did the right thing – because it just happened automatically, and there was no temptation to do otherwise. Also, cultivating virtue might itself require struggle, which might itself add value to it. All of this could be praiseworthy, which certainly sounds like a worthy and valuable thing, but perhaps it would not be if there was no choice – so we would miss out on the value of praiseworthy things.

This is related to autonomy in a way. Cultivating a virtue is (apparently) something that comes from yourself, and so is making the right choice. If you do not have that sort of autonomy, then the virtue or the accomplishment is not yours in some sense. It was just handed to you. And perhaps it is not really a virtue or an accomplishment at all; just a state of affairs without that kind of value.

In summary, many of the reasons why free will is seen as valuable and necessary have to do with autonomy and being the originator of one's own actions – although just what this means is left vague at this stage.[63] This is seen as valuable in its own right as well as allowing ownership of one's own achievements. Morally valuable things also seem to depend on their very existence on real authorship – according to this line of thought, they cannot even exist in a world where people are not free. Further, free will seems to have something to do with rationality and thus

[62]Particularly at least sections 3.3.4, 4.3, 5.4, 8.1, 8.3, and 9.2.1.

[63]Questions related to this are discussed later in sections 2.4.1, 3.4.2, 3.6, 4.2, 4.3, 8.3, 9.1.2 and 9.6.1.

with successfully acting in accordance with one's own values and motives.

These values should be kept in mind later when different conceptions of free will are evaluated. For now, we will move on to quickly surveying the vast terrain of arguments about whether free will is compatible with determinism or not.

2.4 Arguments for and against incompatibility: A quick overview

The question of whether free will should be considered as being compatible with determinism or not has taken many turns and different paths. Those relevant for my thesis will be considered in detail below, especially in chapter 3 below. To give an outline of the terrain, this section presents several arguments and counterarguments for both sides in a very condensed form; some will be discussed later, others will not.

There are two related main threads for incompatibilist arguments. They are mainly based around requirements for causal control and alternative possibilities. Thus, arguments in the first thread cluster around questioning the agent's causal control when events and things outside and before the agent and the agent's choices determine the agent's choices. The second is based on the even more obvious point that determinism does not allow for multiple possible futures and thus, in at least one clear sense, does not allow alternative possibilities. Incompatibilism based on causal control has been termed *source incompatibilism*, and incompatibilism based on alternative possibilities *leeway incompatibilism*.

2.4.1 Arguments related to causal control (source incompatibilism)

Aside from determinism proper, but related and sometimes inextricable, it can also seem threatening to freedom when human actions and choices are explained causally, scientifically or mechanistically, so that agency, the self, souls, or free choice are not mentioned in the explanation. This kind of explanation seems to many to exclude the agent's free choice; everything is reduced to the workings of mindless entities like

genes, brain cells, culture, laws of nature and so on.[64] At least some libertarian conceptions of free will can be seen as the opposite of this, putting the agent and their freedom back in. I will have much to say on (and against) this argument later on,[65] so I will not go into compatibilist counterarguments for this point here.

Closely related to this point and overlapping with is the argument that universal determinism implies that our choices are determined by (prior) factors outside of our control. This is particularly embodied in Peter van Inwagen's consequence argument, which simply put goes as follows

> If determinism is true, then our acts are the consequence of laws of nature and events in the remote past. But it's not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us.[66]

What this means exactly is a more complicated question. Daniel Speak explains the conclusion as follows:

> The argument is supposed to show that the truth of determinism would undermine free will. That is, it is supposed to establish the necessity of the following conditional conclusion: If the way things go in the universe is fixed as a matter of natural law by initial conditions, then no one can ever do anything other than what he or she in fact does.[67]

[64]For more discussion of this idea and related ones, see Knobe & Nichols 2011; Earman 1986, pp. 239–250; Harari 2011, p. 263; Harari 2018, pp. 299–302; Willmot 2016, pp. 7–8.

[65]See 3.6, 9.1.2, 10.7.

[66]Van Inwagen 1983, p. 56.

[67]Speak 2011, p. 116.

This could be criticised for begging the question. If compatibilism is true, then it is hardly true that the kind of lack of alternatives implied by determinism would undermine free will, because that is just a statement of the thesis of incompatibilism. However, this does not quite exhaust the consequences of the consequence argument. For a start, its conclusion that *it is not up to us* or *we have no control over* what we do does not have to be interpreted as saying we *could not do otherwise* in the sense that it is obvious we cannot under determinism. I give a rebuttal of the consequence argument based on my own arguments in 4.1.

Though the arguments presented here share a common theme of being related to the requirement of causal control for free will, they are subtly different from each other, and they are not are purely about causal control. This is best seen with how the first one, which is as much about a threat to intentional control as causal control, and which is also not completely about determinism as such.

Moving on with other related arguments, if everything we do and choose comes down to previous events and we have no agency to choose beyond their influence, the conclusion looms that everything we could ever take of give credit for or blame someone for is merely a matter of luck. Then again, the same sort of objection can be raised against indeterminism. These themes have been mentioned before and will be discussed much below.[68]

Yet another related incompatibilist requirement for free will is that the free agent should be the ultimate origin of their own choices. This would avoid anything else being a more ultimate origin, if that compromises freedom. Two main counterarguments to this are that the agent can be the origin of their own choices even if they are not the ultimate origin, and that is enough; and that being the ultimate origin of anything leads to an infinite regression and/or is self-contradictory. The idea of ultimate origination and the latter argument above against it will be discussed

[68]The importance of praise and blame is described in 2.3. The way it all seems to vanish is discussed at more length in 1.1.3, 5.3. The notion that indeterminism leads to everything being a matter of luck is discussed especially in chapter 3.

in more detail in section 3.6.

Arguments for incompatibility from causal control can also be applied without going back beyond the present moment. It can be argued that if everything the agent chooses follows from their desires, other mental states and so on, the agent themselves has no control, but rather these states do.[69] This is discussed under "homunculus" explanations in 10.7. To repeat the point from that section shortly, if it is a part of you, then it is a part of *you*, and you are not some part of yourself standing apart from all your actual parts. On the contrary to demanding homunculus explanations, I refuse to accept them. Yet, we may not be determined by our desires in a simple manner; see 5.4 and 8.3.2.

One more type of incompatibilist argument roughly related to causal control is the manipulation argument.[70] These arguments start by positing a situation in which the agent is being manipulated by another one outside it and seems to be unfree and not responsible, and then go on to compare this to the situation where the agent is simply deterministic, claiming there is no difference between the two situations. While I do not address or rebut manipulation arguments comprehensively in any single section in this work, different parts of the larger picture I paint are related to them in different ways.[71]

---

[69]See e.g. Reid 2010, Essay IV chapter IV; Visala 2018, p. 121.

[70]See De Marco & Cyr 2024a for an overview on these.

[71]The main counter against manipulation arguments that can be derived from my theory is derived is based on the idea that I present later that holding someone (role-)responsible only makes sense when the person is reasons-responsive in such a way that they can react to being held responsible (8.2, 9.3). If there is a person who is manipulating someone, the responsibility needs to be on the manipulator, because they are the ones who can make a difference to the outcome and thus the ones who can be incentivised to do so, while the victim is unable to affect their manipulation (cf. also 9.5). If there is instead a natural or random occurrence that alters a person's disposition without intention, this mechanism is also not reasons-responsive. Merely changing *inclinations* do not necessarily erase reasons-responsiveness, and an altered person might be partly responsible regardless, but they would still find themselves unprepared in an abnormal situation that probably cannot be treated the same way as business as usual. The idea of some things being counted as part of the person's "real self" and others not also points to a reason why manipulation cases may differ from

2.4.2 Arguments related to alternative possibilities (leeway incompatibilism)

Whenever determinism holds on a level of description, then it is true that on that level, it is only possible that one thing happen next. This also applies when a choice is being made and the thing that happens next is the choice that is being made. More specifically, since (in)compatibilism normally concerns (in)determinism on the ultimate level, if determinism holds on the ultimate level, then any choice can have only one outcome as described on the ultimate level.

The arguments for incompatibilism relating to alternative possibilities are based on the idea that a free agent should be able to do otherwise than they do, and that this means it should be possible on the ultimate level that they do. This sounds initially plausible. The principle that free will requires being able to do otherwise

---

the normal case, though my reductive way of answering the real-self-question in 8.3 does not give very strong tools for answering manipulation arguments, so I would instead focus on the other points before this. (See De Marco & Cyr 2024b for an overview on discussions on the role of the manipulator as a counterargument to manipulation arguments, and De Marco 2025 similarly for discussions on bypassing the agent's normal processes.)

The three modes of thinking described in 5.3 can be helpful for explaining why making comparisons between different situations can make it feel like the normal situation is not free, since such examples may shift one from thinking in the mode of agency to the mode of causality. To the extent that manipulation arguments are about invoking the feeling that a normal situation is like a manipulation case, see chapters 3 and 4 for arguments for why merely evoking the feeling is not enough to prove there is a problem if there is no concrete problem in the normal case. On this last point, cf. also the argument from the literature that either the manipulation makes no difference or there is no worrying analogy to normal cases in the first place (see De Marco & Cyr 2024a, section 7).

On the side of things where I am more in agreement with the point of manipulation arguments, they can be used to demonstrate the idea that nobody can be the ultimate origin of their own choices, since being born into the world is not in all ways that different from, say, waking up one day having been altered. Note that this is not helped by indeterminism. The idea of ultimate origination is discussed in 3.6; see also 1.1.3 for how both determinism and indeterminism lead to similar problems. Chapters 3 and 4 and section 9.1.2 discuss why neither indeterministic "ultimate origination" nor more genuine (but impossible) ultimate origination is needed. However, the point demonstrated is not entirely invalid, and in some sense, it is meaningfully true. We do also need to be able to remember the perspective that nothing is ultimately up to us. This point is touched on in (again) 9.1.2, and a little in 8.2.2.

can be called the *principle of alternative possibilities*, *PAP* for short. The same requirement can also be used as a criterion for responsibility for an act.

Something worth noting in passing is that it typically seems intuitively clear that not only does responsibility require free will, and free will requires being able to do otherwise, but responsibility also requires being able to do otherwise in a direct sense. We apparently do not need to go through thinking about free will as a requirement to come to this conclusion.[72]

One thread of this argumentation is that, in order for an agent to make choices at all and/or deliberate between options, those options must be open to the agent. This could be interpreted in terms of indeterminism. I can already say it must imply indeterminism on *some* level, as the denial of this idea is describing *some* level of description as deterministic in saying there are no options. However, just saying this much is trivial, and just what levels the indeterminism needs to be on is its own discussion. This option is discussed further through specific philosophers in sections 3.4.9 and 5.4.4. In any case, this question is closely related to the question of just what it means to be able to do otherwise.[73]

One compatibilist strategy against the incompatibilist argument from (not) being able to do otherwise is to assert that you could, in fact, do otherwise even under determinism. This is likely to rely on an analysis of *ability to do* that is different from an analysis of *what just is possible*, since determinism does imply that only one thing can possibly happen. Similar arguments have been advanced to say that moral responsibility does not require alternative possibilities, even if free will may require them – famously in Harry Frankfurt's examples that aim to show there are possible

[72]See e.g. Haji 2011, p. 289, Wiggins 1973, p. 48.

[73]On different analyses of alternative possibilities or being able to do otherwise, from different perspectives, see e.g. Wiggins 1973, pp. 48–50, Kenny 1973, p. 102, Honderich 1973, pp. 202–204, Berofsky 2011, Earman 1986, p. 244. There is much nuance here that I am glossing over because it makes little difference for present purposes which exact form of ability or possibility is being (re-)interpreted, as long as the result is that it ends up being compatible with determinism.

circumstances in which an agent is responsible for acting from their own motives even though they could not have done otherwise if they did try.[74]

One version of the alternative analyses of being able to do otherwise is the voluntarism of classical compatibilism, which states that to have free will is just to do what you want to do. I will not discuss this idea much other than sometimes to point out how it does not work and something more is needed in a particular context. Shortly put, it seems to be more about the lack of outer constraint, not about inner freedom, and it cannot address cases like addiction or irresistible desires, which are usually considered unfree, and which I also consider so.[75]

Christian List lists three kinds of analyses of what it means for an agent to have "alternative possibilities":

> **Conditional interpretation:** If the agent were to try or choose to do otherwise, he or she would succeed.
> **Dispositional interpretation:** The agent has the disposition to do otherwise when, in appropriate circumstances, he or she tries to do otherwise.
> **Modal interpretation:** It is possible (in a sense to be spelt out further) for the agent to do otherwise.[76]

Note how even the modal interpretation still contains the ambiguity of just what is meant by possibility.

One analysis of how alternative possibilities can also exist under determinism is given by Christopher Taylor and Daniel Dennett[77]. It shows an example of a fairly

[74]Frankfurt 1969.

[75]See e.g. Honderich 2002, pp. 105–108. In the present work, section 8.3.1 is related to the insufficiency of voluntarism.

[76]List 2019b, p. 81. Bolding in the original. Formatting not reproduced exactly.

[77]Taylor & Dennett 2011.

technical argument appealing to things such as analysis of causality and possible worlds, but one simple point that can be explained here quickly as an example is that in other kinds of cases, when we speak of what was possible in a situation, we do not assume the possible worlds examined needed to be *exactly* the same up to the point where the different possibility was realised. If so, why would it need to be so when we are talking about PAP for free will? Merely similar worlds can have different futures under determinism.

Oddly, Dennett has also given an argument that PAP does not matter.[78] In his pointedly named piece "I Could Not Have Done Otherwise – So What?"[79], he questions whether there is any reason to regard this principle as worthwhile, and states that he finds little[80, 81]. His idea is that if responsibility does not require alternative possibilities, neither does freedom. For one thing, he writes, it is not true that ordinary people are concerned with being able to do otherwise in some deep metaphysical sense; their interest in whether someone had the ability to do otherwise in the sense of not being under what he calls local fatalism, i.e. that one cannot change things no matter what they do, say because they are locked in a room and cannot get out to do what would be necessary[82]. Secondly, he points out that there are examples of not being able to do otherwise that do not diminish responsibility. If Martin Luther said that "he can do no other" because he could not fail to follow his principles, is he not strong of will rather than weak, and all the more responsible for

[78]This is not to say that his claims in the two articles contradict each other if one understands in what sense he means "being able to do otherwise" in each case.

[79]Dennett 2002.

[80]Dennett 2002, p. 84.

[81]Note that the self-proclaimed libertarian David Peroutka (2022), mentioned below at 3.2.3, proposes a similar idea, though I do not look at his version closely in that sense.

[82]Dennett 2002, p. 85.

standing behind his principles?[83] A third point is that if the question really about whether a person could have done otherwise under those exact circumstances (see 11.8, and also compare 5.4), then we would be in the peculiar situation that, given the current state of physics, we would never really know whether anyone has been responsible for any act, since we cannot really tell how deterministic our brains are.[84] This point along with other similar ones is touched upon again in 2.5 shortly.

Arguments like these have prompted incompatibilists to state that they are sure there are kinds of freedom that are compatible with determinism, but those are not what the incompatibilists are talking about.[85] Of course, this is not really in agreement with what compatibilists mean; they will want to say that the kind of freedom they speak of is the real thing (cf. 3.1). Still, it is certainly true by our definitions (see 2.1, Appendix A and Appendix B) that universal determinism on the ultimate level of description makes alternative possibilities nomologically impossible on the ultimate level. *If* that is enough to preclude free will, as libertarians in the sense used here (see 1.3) hold, then there is no getting around that contradiction between freedom and determinism.

In sum, determinism makes alternative possibilities impossible in some sense but leaves them possible in some other senses. Instead of arguing about whether this "really" means they are possible, the question that we need to ask is which of these senses actually matter to free will, and why – or whether any of them do. If incompatibilists say they want the possibility of other alternatives in some specific sense, then it is not an answer for compatibilists to say they are possible in some different sense. However, we should definitely be asking what the reasons are for wanting alternatives in some sense or other – and whether those reasons are good.

An important point that rises from this discussion is the following: if being

[83]Dennett 2002, p. 86.

[84]Dennett 2002, pp. 87–88. See also Dennett 2015, chapter 6.

[85]E.g. Searle 2007, p. 47, Kane 2002, p. 223. See also section 6.2 in the present work.

able to do otherwise is only possible under indeterminism, then "is able to do otherwise" implies "may actually do otherwise in the exact same circumstances." This might seem innocuous, but it gets libertarians into trouble later, forming the basis of the randomness argument (see 3.3). It might also seem trivially obvious, but it is surprisingly easy to forget in practice (see 2.2.2 and 3.2.1).

2.5 Some oddities in thinking about free will and determinism

There is a set of arguments and observations about free will and determinism that goes both ways that does not really fit under any other heading but needs to be discussed. These points concern how odd it is to think of freedom and determination at the same time, whether assuming incompatibilism or not.

There are various strands of argument against determinism itself in the sense that it could not be consistently held, perhaps contradicting the very idea of the person holding it thinking or reasoning. Some of these have more to do with antireductionism than determination. I will only focus on such ideas here as they relate to choice and decisions specifically.[86] Another specific argument of this family is discussed later in 3.4.9.

There seems to be something strange about the very idea of taking determinism as a starting point when making decisions. Charles Hartshorne, arguing against decisions being determined, writes as follows:

> Isaiah Berlin reports J. S. Austin as saying that while many talk about determinism, no one really believes it, "as we all believe that we shall die." This latter belief we take into account in our decision making (not enough, to be sure, but still we do take it into account). In contrast, it is impossible to take determinism into account, for it has no consistent practical meaning. *Before* I decide I may claim to know that my decision will be fully determined, whether by heredity or environment, or by God,

[86]For an overview of some such arguments, see Honderich 2002, chapter 7.

> but in what way can my decision take this alleged knowledge into account? *After* the decision I can say, See what I was preprogrammed to decide! But this in no way or degree helped me to make the decision. It was an idle retrospective application of a useless doctrine. The application in decision making is always too late.[87]

No doubt Hartshorne as an incompatibilist thinks this shows how absurd the notion of determinism is in the context of decision-making. It can also be taken in the converse way. Incompatibilists are the ones who think that something special should follow from determinism being true of our choices. However, if we imagine making a choice while determined, then from this kind of a point of view, it seems nothing at all follows beyond what choice-making is already like. Compatibilists can thus say this shows nothing bad follows from determinism that contradicts free will.[88]

If you were to think of yourself as determined in such a way that you are not making any choices yourself, that, if it is even coherent, might have more dramatic results. I argued this back in my Bachelor's thesis: If you were to decide to forgo making decisions and instead wait passively for the causal chains of nature – which you would see as separated from any hypothetical choices you might have been thought to make – to move you, then you would *not* be moved to do anything as long as you were acting like this. After all, even if it was a causal chain like that that causes all your putative decisions, it would currently be causing you to make no further active decisions. There would not be a causal chain from outside of you making you do anything; just the one inside you, and if you were taking the option not to do anything for internal reasons, there would be nothing else doing it for you. You would have to treat yourself as capable of taking one option or other out of those

[87]Hartshorne 1984, p.19. Italics in original.

[88]For ideas in the lines that it is harmless to think of actions as determined, see e.g. Dennett 2015, pp. 114, 118, 138–139, 169; Harris 2013, pp. 36–37; Hart 2008, pp. 46–47.

available to you in order to be doing anything.[89]

However, this is not strictly speaking true. Consider Susan Blackmore's description of her own experience:

> I used to have two possible routes home, the main road and the prettier but slower lanes. As I drove up to the junction I was often torn by indecisiveness. … One day I realised that 'I' didn't have to decide. I sat there, paying attention. The lights changed, a foot pressed the peddle [sic], a hand changed gear, and the choice was made. I certainly never went straight on into the stone wall or banged into another car.[90]

There is nothing logically impossible about things happening this way, and analysing what could be happening there shows the logical loophole in the argument that thinking in terms of not being in control precludes making choices: If the conscious part of the mind lets go of making decisions (let us say on a particular occasion), then decisions can still get made by the person if the unconscious part of their mind is capable of doing it instead.

Of course, saying a person's unconscious mind is making the decisions and their conscious mind is not sounds a lot like saying the person is not really making free decisions. That is indeed something Blackmore thinks[91], and others such Sam Harris say something like this as well[92]. This topic is discussed in a little more detail in section 9.6.1, though as explained there, the question of consciousness is strictly speaking outside the scope of this work. For now, we can just observe that, in a statement that may be an oversimplification but is roughly true, refusing to think one

[89]Kokko 2011, p. 25.

[90]Blackmore 1999, p. 244.

[91]Blackmore 1999, pp. 236–237.

[92]Harris 2013.

has control makes it impossible to consistently make conscious decisions, but not for some kind of decisions to get made, whatever the status of such unconscious decisions may be.

A final related argument about determinism and free will, this one on the compatibilist side, is that it would be absurd if we needed to know whether the universe is deterministic or not to know what the meaning of our actions is. The world could have appeared exactly the same way in every respect, but then, at some point in the future, we would find out a deep truth of physics that the world is really deterministic at the bottom, and that would force us to think differently about all our human-level actions and relationships. There certainly seems to be something off with that picture.[93]

2.6 What determinism does not threaten

In this section, I will deal with some objections to compatibilism that are apart from the main argument and can, I think, mostly be dismissed as misunderstandings. These are the question of whether determinism implies or is similar to irresistible desires that limit freedom; the related questions of "special determinisms" such as psychological or genetic determinism; the question of determinism implying inevitability; and the question of whether determinism threatens creativity.

There is no hard line to be drawn between which topics belong here, to be dismissed as misunderstandings, and which should be discussed in chapter 3 as potential threats to freedom from determinism even if they are then dismissed, but this rough distinction serves to keep that chapter from growing even more bloated.

2.6.1 Irresistible desires and "special determinisms"

One way determinism might seem to be contrary to freedom is if we consider a case where someone is determined by a particular desire to act upon that desire, no matter

[93]Cf. Dennett 2015, pp. 148–149, List 2019, pp. 158–159.

what else. The desire could come from a compulsion or an addiction, and such cases are generally seen as unfree, not only by incompatibilists but almost everyone. One might go further to ask what the difference is between being determined by such compulsive, unfree desires, and the general idea of a person being determined by their desires.

A quick initial answer to this is that a person being determined by a single, compulsive desire is different from someone who can consider which action to take and which motive to follow, weighing them against each other. More on this point can be found in 11.8 and 5.4. In terms of this section, the explanation of the point overlaps so much with the two following sections that I refer the reader to them.

Another related question that should be kept from determinism proper is the question of what Peter van Inwagen calls in one place *special determinisms* – as he continues, "that is, theories that say, or are sometimes interpreted as saying, that some important aspect of human behaviour determined by this or that factor outside our control[.]"[94] His example just there are claims that sociobiology – the modern equivalent to which is evolutionary psychology – shows that women's role in society is genetically fixed and cannot be altered.[95] In general, the reasoning here applies to any claim that humans are determined to always behave in some way, because we are programmed to do so by our genes, or by other psychological or evolutionary reasons, or something else, even social media algorithms[96]. I will not discuss the specific claims much, as the point is only to show why the question raised by them is not the same thing as the compatibility question in general.[97]

---

[94]Van Inwagen 1983, p. 100.

[95]Van Inwagen 1983, p. 99.

[96]On this last, see Lanier 2018, p. 28. Incidentally, this source implicitly takes a compatibilist stance.

[97]Richard Dawkins in 1999, chapter 2 offers a noteworthy rebuttal of the idea of genetic determinism, relating it somewhat to the question of determinism in general.
There is an irony in claiming that since science uses such reliable methodology, scientific findings about what "determines" human behaviour should be taken as overriding.

The core point here is simple: the incompatibilist claim is that determinism in general contradicts free will. Some kind of special determinism, if it were true, might threaten free will even if this were not true. The difference between a special determinism and just determinism is that under mere universal determinism, many kinds of factors could go into determining our decisions, whereas a special determinism would override other kinds of influences, such as our desires or reason. That could very well be threatening to free will regardless of being determinism[98].

As a further demonstration of the difference between determinism and the kind of influences thought of as potential "special determinisms", the special determinisms need not even be literally deterministic to be a problem. If they were deterministic, that would be a question of higher-level determinism, but even that does not need to be the case. Taking the example of gender roles being "determined" from above, suppose that it was true in almost all cases that all societies must have the same sort of gender roles, but it would be possible to change this in some rare and demanding conditions, such as every person being given an inbuilt AI-assistant that would enable them to compensate for the limits of their biologically determined, sex-based thinking abilities. This would not be deterministic on the level of description where you simply speak of the unalterability of gender roles, but that does not mean it would not be problematic the same way. In another example, it

---

This is because of what that methodology involves, in order to be reliable: eliminating confounding factors. To be reliable, a psychological (etc.) study needs to have a large enough sample size, randomised control and experimental groups and so on to eliminate random variation and the effect of other factors than the one being studied. Studies also need to be repeated before the results can be trusted enough. (See e.g. Stanovich 2003.) Results found with such methodology may indeed be reliable in detecting real causes because they used such methodology – but to claim they therefore show features of behaviour that are unavoidable or cannot be overridden by other factors is to get things exactly backwards. The thing that makes such studies reliable is their eliminating all the *other factors* that can and do have an effect. If the factor found at the end had been determining by itself, there would have been no confounding factors to eliminate.

[98]See also 5.4.

would be similarly problematic if the claim was that gender roles are not literally impossible to change, but doing so will lead to the collapse of society.

What this means is that just because special determinisms would threaten free will does not mean that determinism in general would do that. Thus, the question of special determinisms is not part of that discussion. Now, discoveries in science may in fact threaten the idea that we have free will. I will discuss this in 9.6.1 against the context of having finally offered a definition of free will.

2.6.2 Determinism and inevitability

One way to look at determinism is to say that, under it, what happens happens "inevitably". This can also be used as a way of articulating why determinism is supposed to be a threat to freedom. However, as Dennett has argued, this idea is problematic.[99]

Inevitable means unavoidable. Is it true that, under determinism, things happen unavoidably?

Suppose that determinism holds and it looks as if something bad is going to happen – at least, it will happen unless there is an agent who can stop it from happening. So maybe a baseball is flying on a trajectory towards Alice's head, or Bob is feeling an inclination to drink alcohol before going to work. Under determinism, is it possible that Alice or Bob will avoid the unpleasant effect? Of course it is. It may be that the person who wishes to avoid what would have happened is aware that it would happen if they did not do something – and can and does something to avoid it. So Alice could see the ball coming and duck to avoid it, with her ducking deterministically explained by her seeing the ball. This possibility has absolutely nothing to do with a denial of universal determinism. (See also section

[99]This section's argument is based on Dennett 2003, pp. 56–62, though with different examples. (Thus, when I say that something "seems to me" to be in a certain way, it is implied that this is both how it seems to me and in accordance with what Dennett writes.) See also Dennett 2015, pp. 114–117 and 134–142.

11.8.)

So, there is a sense in which things can be “evitable” (avoidable) under determinism. Whether things are avoidable in such a sense also bears on free will. Suppose Bob is a caricature of an addicted person and cannot resist the urge to drink in spite of knowing his boss will notice the consequences at work and might fire him – an example of an irresistible desire. This is a basic example of a case where someone might be said not to be free. Its converse is, again, possible under determinism: Suppose that, tempted though he is to drink, Bob considers the consequences and decides not to, because it would go against more important interests of his. In this case, his inclination to heed those interests can perfectly well deterministically lead to his abstaining from drinking.

Alice could be too slow to dodge the ball, in which case having it hit her head would not be avoidable for her. This would be a case of limits of (another kind of?) ability, however, rather than lack of freedom. Meanwhile, if Bob was for internal reasons unable to refrain from drinking, this could be seen as a limit of his freedom itself. In neither case is the difference between avoidability and inevitability due to universal determinism. Now, if Bob were to be so stereotypically addicted that there was a deterministic (or even almost deterministic) rule governing his behaviour that stated he will always drink under certain kinds of circumstances no matter what, that kind of determinism would lead to inevitability – but that is a different question. It is also an idea that will be very important later (see 5.4), but for now, it will be set aside.

That is one sense of “inevitability”, and a relevant one. Yet, I freely confess that there seems to be some other sense in which it is true that determinism implies “inevitability”. If determinism holds, then whatever happens could not have happened otherwise, at least in some strict sense, so does that not fit the general idea of inevitability? Sure, a creature (or just an inanimate thing or event) could have prevented something from happening that would have happened if they had not interfered – but whether they did stop it or not could also not have been otherwise.

What is this other sense of inevitability? How shall we characterise it? All I can think of is that it means the *same thing* as determinism. I agree that I feel we are using the word correctly if we say that it was "inevitable" for something to happen when, given that things before that were exactly as they were, it could only have happened that way. If we do, though, we are only using the same definition as for determinism, not talking about something further that follows from determinism. Thus, it does not give any reason to fear determinism. To fear that kind of "inevitability" is just to fear determinism, and if there are reasons for the fear, they have to be explained separately. We have reason to fear being like Bob with his stereotypical addiction (or Alice who was too slow to dodge, for that matter), but that kind of unavoidability is independent of determinism.

Another way of looking at this is that "inevitability" could mean either determinism or fatalism. Determinism means things are determined in a forward-looking (roughly causal) way, where what is now causes what will be alter. Fatalism means things are determined in a teleological manner, meaning that what will happen later cannot be altered beforehand; you might do different things, but they will all lead to the same outcome.[100, 101]

Perhaps there is some potential for confusion, though? Perhaps, since "inevitability" somehow seems to refer to both mere determinism and actual unavoidability, we are somehow prone to confuse these two ideas? This guess proves

[100]Thanks to Eeva Tolonen for this insight.

[101]Depending on how fatalism is understood, it may be a stronger or weaker thesis than determinism. Fatalism with respect to some event *X* means that *X* is determined to happen. If we say merely this, then determinism is a stronger thesis, because it says that not only *X* but everything else is determined to happen. However, fatalism can also be understood as saying that *X* will happen no matter what else happens before the time *X* happens. In this case, it is a stronger thesis than determinism in one way but still weaker in another: stronger because determinism would allow that only some chains of events (those derived from some initial conditions) lead to *X* and others do not, though still stronger in the sense that fatalism could allow that things other than *X* could possibly happen in more than one way.

nothing in itself, but I will return to such themes later (chapter 5).[102, 103]

2.6.3 Emergence and creativity

One possible argument that needs to be addressed (if briefly) is that a deterministic universe where everything is "determined in advance" would threaten the possibility of creativity and novelty. Though there is a kind of simple logic behind this – that what is "determined in advance" cannot be "novel" – it is a strangely subjective one, nevertheless.

It is not true that a vast universe that is deterministic will produce nothing new. The scientific world view suggests, without requiring indeterminism for this suggestion, that the universe has constantly been evolving new phenomena during its existence. Further, the possibilities of combining the existing elements are so vast that there is no time during the lifetime of the universe to explore them all.[104] Whether things are deterministic or not, new things as in things that have not existed before can emerge, and we can see this happens and can happen in a very interesting way. This is a very abstract argument on a level of description far removed from the human aspect of creativity, but this simply follows from the idea that the threat is attributed to determinism on the ultimate level.

---

[102]Denyer 1981, p. 50 argues that determinism implies "fatalism" – much the same thing as *inevitability* here – because determinism makes the future necessary and only the contingent can be subject to practical deliberation. Though this is answered in the present section in some sense, it is worth also looking at 3.4.9 for a more thorough answer to this sort of argument.

[103]Cf. Slattery 2014, pp. 220–221, answering to an unidentified philosopher who claims that determinism would mean there is no point in trying to communicate anything to people if the world is deterministic. Slattery here makes a point about how determinism is not inevitability in this specific context, although he has stated (personal correspondence, or more specifically a public online conversation) that he thinks Dennett's argument that I followed above to be fallacious. I do not know what he would say the difference is.

[104]See e.g. Kauffman, 2008.

In section 2.3, I mentioned the value of creativity as a possible reason to want free will. This section partly addresses those concerns, but note that the question posed back there was more about being the origin, the true creator of your creations. This section is addressing the question of it being possible to bring about something novel, and thus, the question of being able to be the creator of something is not addressed here in all ways. For that, see 4.3.2 and 9.1.2.

All of this said, I think the value of spontaneity, novelty, creativity and so on is a topic and perspective that would merit exploring in its own right. There is probably more to be said than just that technically, this does not threaten compatibilism. However, in the interests of space and sticking to the main thread of my argument, I only say that here.[105]

2.7 Conclusion to chapter 2

In this chapter, we have quickly reviewed several of the multiple branches of the determinism/free will debate. We have seen that much of it has to do with whether various requirements mostly related to alternative possibilities and being the origin of one's own actions are compatible with determinism. We have also seen that much of the debate seems to have to do with contradicting intuitions, even contradictions within the same intuitions, and seen hints that there may be two kinds of conceptions of free will, some of which are incompatibilist and some of which are compatibilist.

Since compatibilists may be accused on equivocating on this point[106], let me reiterate this explicitly: determinism means that, in the strictest sense, on the ultimate level, it could not have happened that you would have chosen otherwise than you did in any individual situation. Compatibilism thus may have to offer other kinds of possibilities of doing otherwise, or something else than the ability to do otherwise, but never alternatives on the ultimate level in the exact same situation.

[105]Something a little related is found in section 3.5.

[106]See, e.g., the quote at the beginning of 3.1 and the whole section 3.4.1.

This might sound like a problem for compatibilism. However, after we see what kind of problems being “able” to do otherwise in the strictest sense causes for the libertarian, we may see it as more of a victory to avoid it. It is to these problems, explained in the randomness argument, that we turn in the next chapter.

3 The Randomness Argument against Libertarianism

Though this chapter is not written as a response to Robert Kane specifically, it can be framed around answering to two different claims made by Kane. The first is summed up by this truncated quote:

> [W]e can make sense of an incompatibilist view of free will and responsibility without reducing it to mere chance or mystery[.][107]

The other is this:

> Many kinds of freedom worth wanting are indeed compatible with determinism. What we incompatibilists should be insisting upon instead is that there is *at least one* kind of freedom worth wanting that is incompatible with determinism.[108]

This wording brings to mind the subtitle of Daniel C. Dennett's book *Elbow Room*,[109] "Varieties of Free Will Worth Wanting", which expresses the idea of the book very well.[110] Why would you fret about determinism if indeterminism is not

[107]Kane 2017, p. 2479.

[108]Kane 2002a, p. 223, italics in original.

[109]Dennett 2015.

[110]Perhaps better than all the rest of it, if you agree with those who think that the book's details leave its overall point obscure (which I have heard or read in various informal conversations; I perhaps do agree).

required for any conception of freedom that is worth wanting? Kane counters that it is.[111]

Thus, opposing these two, I aim to show two things: Firstly, indeterminism does always mean "randomness", at least in some relevant sense. A model of free will involving indeterminism need not be utterly random, but indeterminism always adds randomness if it adds anything, and this matters. Secondly, there *is* no form of necessarily indeterministic freedom worth wanting, at least insofar as it is not only worth wanting for the circular reason that one wants to have indeterminism itself. These are the two conclusions of what I call the ***randomness argument***, which is the main concern of this chapter.[112] The discussion of the second question will also continue in the following chapter. Before I introduce the randomness argument, I need to go through some preliminaries.

For a condensed summary of the entire argument and its various sub-arguments, see Appendix C.

[111]Compare with Balaguer 2014, pp. 52–54, where the author merely asserts that he only cares about a kind of incompatibilist free will, not for some reason but as a starting point. While this is as it were valid, in that anyone can make a choice like that, it is premature to do when there is a possibility to examine the desirability of different options more closely, maybe discovering things about them you did not realise before. In any case, if one does make such a choice without further reasons, it is no argument why others should do the same.

[112]A more condensed version of the argument of this chapter can be found in Kokko 2024c.

3.1 Why ask for the “right” definition of freedom?

> If the compatibilist defines free will in a less common way, and in a way not in conflict with determinism, (or indeterminism, for that matter), the important part is that they be true in regards to the implications of their definition, and they make their definition so crystal clear that no person with a different (and more common) definition could confuse it.
>
> -’Trick Slattery[113]

A question someone might ask at this point is this[114]: There are different definitions that may be given for “freedom” or “freedom of the will”. Is it at all interesting to debate about which definition is “right”? The concept is surely ambiguous, and it might be better to acknowledge the ambiguity and move on. Why not just say what definition of freedom I am using and continue talking about that while letting the other conceptions alone?

That general point is often a good one. There is no value, in itself, in arguing about what definition of a term is the “right” one. If we are to talk about what is “real” free will, that question must have some meaningful sense, and that sense should be stated. What is it in this case?

Suppose that I did say that I am going to speak of *free will in a certain compatibilist sense*; that when I write “free will”, the reader is to understand it in this sense. Then I would go on to explain the rest of my thoughts on free will and the human condition, as I will in later chapters. If a libertarian or other incompatibilist were to read this, what would they have reason to think? Fine, the libertarian might say, go ahead and talk about that sense of “freedom”. We will see what follows from it. It may be an interesting exercise. It will not, however, tell me much about the

[113]Slattery 2014, p. 334.

[114]Thanks to Veli Virmajoki for asking it. Compare also Balaguer 2014, pp. 49–54 and Vargas 2013, pp. 11–12.

human condition. What you describe is not how I think humans are, and what you call freedom is not the thing that I desire when I say I want to be free. It is something else – something hypothetical and not important or meaningful.[115]

In other words, if I were to just talk about "free will" in some particular sense of the expression, without justifying my choice as the "right" one, I would not have justified my claim that I am saying something meaningful and interesting about the human condition. What I mean when I say I want to find the real sense of freedom is that I want to see what sense of freedom is meaningful to us (assuming there is only one), what it is desirable that we should have. Insofar as the reader accepts this argument, they will then have reason to think that, with respect to the question of determinism and indeterminism, I have made the correct choice to go on talking meaningfully about free will and the human condition.[116]

3.2 Logically possible options with respect to determinism

The stage is now set to move onto my argument about what determinism and particularly indeterminism imply for freedom. Recall that I gave my definition for determinism 2.1.4 as the following:

> Determinism holds in a part of the universe P at a level of description L iff, given any total state of P as described in the terms available in L at time T_1, and any total input from the rest of the universe between T_1 and T_2 inclusive, as described in the terms available in L, only one total state of P as described in the terms available in L is nomologically* possible at T_2.

[115]Cf. Donagan 1987, p. 183.

[116]Cf. Dennett 2015, pp. 2–4, 6, and p. 57, footnote 6. See also Nichols 2006, pp. 58–60 on the "descriptive, substantive and prescriptive projects," all of which are undertaken here to some extent, but with emphasis on the prescriptive in the end.

Conversely, indeterminism holds in *P* whenever determinism does not. In the next section, I will apply these definitions in the central part of my argument. To start with, I need to make it clear what they make possible – and what they do not.

3.2.1 Why there is no true third option

Indeterminism is the denial of determinism. As with defining determinism, I am keeping this as simple as I can. If universal determinism does not hold, then indeterminism is true of the world as a whole. Similarly, if determinism does not hold in some limited part of the universe, then indeterminism does.[117]

This point should be understood very clearly. Determinism was defined as only one possible future state of affairs being nomologically* possible given certain starting conditions, and nothing more. Anything that does not fit this definition is, by definition, indeterminism. These definitions make it clear that there is no option that is neither determinism nor indeterminism. Anything whatsoever that is not determinism is indeterminism.

I having made the choice to use these definitions, it now falls on me to use the terms in arguments where such broad definitions are nevertheless enough to secure the conclusions that I arrive at. Thus, if I want to conclude that something follows from indeterminism, I need to be sure that it follows from anything except complete determinism. Before going there, I certainly need to acknowledge the variety of options left outside strict determinism, as well as the variety in what may be available within determinism itself.[118]

3.2.2 Intermediate positions

When we say that there is no true alternative to both determinism and indeterminism

[117]As always, this is relative to the level of description used.

[118]See Slattery 2014, p. 104 for a brief articulation of why manyvalent logic also cannot be applied to escape the dichotomy.

because indeterminism is defined as everything but determinism, it is not hard to guess that there are multiple possible forms of indeterminism that are intermediate between determinism and what we might imagine as "total indeterminism". In addition, there are forms of determinism that resemble indeterminism in relevant ways.[119]

It was already brought up that indeterminism may be confined only to a part of the universe. Just to be clear, this would be defined as the rest of the universe matching the definition of being deterministic. Another possible limit for indeterminism is that the laws of nature are such that an undetermined event can nevertheless only realise one of several predetermined options. For example, in radioactive decay, a particle may decay or not decay at any given moment, but it cannot randomly do just anything, such as multiply into three particles of the same sort or explode with the force of a supernova.[120] Different degrees of freedom for the indeterministic event are possible here. It could (purely hypothetically, not in real physics) be that a particle may move in a completely random direction (while still not being able to triplicate instead), or that it could randomly move in one of two given directions but no others.

Another possible limitation is that an indeterministic event may be probabilistic.[121] The decay of radioactive particles, again, happens at a certain

[119]Cf. Vargas 2013, p. 9.

[120]In fact, any "one thing" in the universe being completely indeterministic without any such bounds would amount to the universe being completely indeterministic and thus completely without structure of any kind. This is because if there was anything that could *not* happen in the universe because of the indeterminacy of the one thing, that would be a limitation for what the thing could do.

[121]Explaining what this exactly means is not necessarily a simple matter, but I will not go into the details here. For an introduction to the concept, see e.g. Niiniluoto 2007. Slattery 2014, pp. 121–124 claims that something being probabilistic can only describe our knowledge of it rather than how it is, and that such things can only be deterministic; this seems to be based on a false dichotomy, but it does highlight some of the problems of the concept.

probability per each interval of time of a given length, depending on the type of particle. This is why radioactive substances can be described as having a half-life: the time in which, from given great quantity of particles, half will have decayed. In other words, when an indeterministic event is dictated by laws of nature to be probabilistic, it has a certain chance of occurring, and given many occasions of that type of indeterministic event, their rate of occurrence will converge towards that rate.

The example of radioactive decay also demonstrates another point: higher-level events composed of large enough numbers of probabilistic events will be virtually deterministic. There is, in practice, a law that radioactive decay will halve the amount of a given radioactive substance in the time indicated by its half-life. More generally, too, it could be that an indeterministic event is very close to being deterministic, or even in practice completely deterministic. Something being "determined" to happen at a 95% probability would in many cases have the same practical consequences as its being absolutely determined to happen, even though technically it would be indeterministic. Then again, given many enough repetitions, it would become likely that the thing sometimes does not happen, which could be significant.[122]

All the previous intermediate positions exist due to the fact that determinism is narrowly defined and everything else falls outside of it. Nevertheless, there are meaningful alternatives within full determinism as well. I will bundle the relevant ones up under the label *pseudorandomness*. The pseudorandom system is one that is in fact deterministic on the lower level, but is so unpredictable on the relevant higher level that it appears indeterministic. Shortly put, pseudorandomness is indeterminism on the relevant higher level (see 10.3) accompanied by determinism on the ultimate level (see 10.6), or at least on some relevant lower level.[123] This can

[122]Cf. Kokko 2014, pp. 75–76. Cf. the endorsement argument (e.g. Almeida & Bernstein 2011, pp. 490–493), which claims that, even with the world being indeterministic, we must have fairly high probabilities of doing what we decide in advance to do, because otherwise, we would not so commonly endorse our decisions.

[123]In general, the phenomena of determinism-like indeterminism and indeterminism-like

take different forms, for example a *chaotic* system that deterministically produces the same results from the same exact starting conditions but is so sensitive to differences in the initial conditions that arbitrarily small variations in the conditions can produce indefinitely large variations in the outcome. Thus, a chaotic system is bound to be indeterministic on a coarse-grained level. In general, pseudorandom systems could appear as if they embody any of the forms of indeterminism described above. The reason they are worth mentioning is that they can be identical to actual indeterministic systems for all practical purposes, and yet, they are technically deterministic on the ultimate (or lower) level. This is naturally the converse of the above point about indeterministic systems that might as well be deterministic.[124, 125]

In the other direction, things could be "more determined" than just deterministic. Something like genetic determinism or determinism based on addiction could be more constraining than being determined only by the exact state of everything. This could be a threat to freedom even if determinism in general is not. These topics are discussed in sections 11.8, 2.6.1, and 5.4.

In sum, though there is strictly speaking no third option between determinism and indeterminism, there are alternatives between a totally rigid determined picture and utter randomness – even utter randomness about only a particular event, such as a particular choice. Further, some forms of determinism have consequences basically identical to those of indeterminism, and some forms of indeterminism have consequences basically identical to those of determinism. I have to take all of this into account as a I present the randomness argument – which claims that all forms of indeterminism can only lead to a lack of control, at least in principle.[126]

---

determinism are related to the possibility of determinism and indeterminism coexisting on different levels that is described in 11.3.

[124]Again, cf. Kokko 2014, pp. 75–76.

[125]Cf. also Slattery 2014, chapter 18.

[126]Slattery 2014, chapter 19 discusses the relationship between different models of time, and it offers an interesting perspective in that from the point of view of models of

3.2.3 "Libertarians" who are not incompatibilists?

Another relevant question raised about intermediate positions is whether some "libertarian" or "incompatibilist" position could be so intermediate that the randomness argument does not apply to it simply because the theory itself does not count as indeterministic or incompatibilistic by my terms.[127] This is always a possibility, and in such a case such a theory will simply not be touched by the randomness argument, but I have not yet come across any actual example of a theory straightforwardly terming itself "libertarian" or "incompatibilist" but not being incompatibilist *at all* in my terms. Thomas Reid's (see 11.9) terminology comes close, but he uses none of these terms, his work presumably predating the current usage. Christian List's "compatibilist libertarianism" also comes close, but as it claims to be compatibilist at the same time, we can see that it cannot be exactly what it says in terms of how the terms are used here (and in general). Besides of this, it still manages to be vulnerable to the randomness argument thanks to its use of higher-level indeterminism (see section 3.4.3).

An example of a position that is remarkably intermediate is David Peroutka's partial compatibilism, which put simply states that wrong acts cannot be free if they

---

time, there *is* an option that is neither deterministic nor indeterministic: the option where time does not exist, since (in)determinism involves change over time. Of course, as Slattery points out, this is no help with any kind of freedom either, since there are no actions to be free about if there is no time. In this work, outside this footnote, I simply assume that time does exist.

That said, I do not have an idea of what it would mean for time not to exist, other than a moment frozen in time that is evidently not how the world really is. Discussion of A- and B-theories of time does not give such meaning either, even though B-theories have been said to mean time does not exist. I think the very distinction between them is a mistake (see Kokko 2024a), and regardless of that, determinism or indeterminism is perfectly explicable in a B-theory. Nothing in my definition of determinism assumes either kind of theory.

[127]It certainly seems to be a common first reaction from libertarians when confronted with the randomness argument to say that what they mean by indeterministic free will is not the kind the argument is talking about – which does not mean they are usually right about this on closer examination. I say this mostly from personal experience rather than specific written sources, though see also 3.4.1.

were determined, but right acts can be.[128] Peroutka has told me in personal communication that this is an example of "libertarianism" that is really compatibilist, and that it is a form of source incompatibilism. Certainly, it is right to describe this view as partially *like* compatibilism and partially not. Peroutka even rightly points out in his article that there could be a world with a different history in which every person only makes deterministic choices, but all choices are free because these are morally right choices. The idea that wrong choices and right choices have different standards also makes some intuitive sense and seemingly resolves some issues, as it might seem more appropriate to praise someone for having unshakeably good motives than to blame someone for having unavoidable bad motives.[129] Limiting the requirement of indeterminism to cases where one actually did wrong even partly disarms the argument I will present later about how the possibility of doing otherwise lessens control. (See especially 3.5.) Still, since indeterminism is sometimes required, this view technically falls on the side of libertarianism in my terms, and it is vulnerable to some form of the criticism that indeterminism is not needed in principle (the second main claim elaborated below).

It seems inevitable that among all self-proclaimed "libertarians" in the world, there must be some who are not incompatibilists. Since I have yet to find a full example, however, this seems to be rare. Such "libertarians" are definitionally not what I call "libertarians" for the purposes of this study, and the randomness argument is not aimed against them. Since seemingly almost all self-proclaimed "libertarians" are libertarians by the terms I use, however, my terminology matches existing use well enough, and I am very much addressing the discussion as it exists.

[128]Peroutka 2022.

[129]This relates indirectly to my eventual argument that free will and morality are related: see 8.3 in the present work.

3.3 The randomness argument

The randomness argument is the focus of this chapter. It says essentially that any possible kinds of choices that are indeterministic, even in part and within limits, are "random" and outside the agent's control, and thus all indeterministic models of freedom face a serious, unavoidable problem. This argument apparently goes back to ancient Greece[130], and forms of it have been fully or partly recognised and utilised by various philosophers, not all of them compatibilists. Meanwhile, others reject it or fail to recognise it.[131] My formulation is meant to be airtight and as directly to the point as possible, leaving no room for attempts to wiggle out of the conclusions by appealing to anything irrelevant.

3.3.1 Earlier formulations of the randomness argument

Versions of the same basic idea that indeterminism contradicts free will because it is something like uncontrolled randomness have been presented numerous times before. Many of these are referenced elsewhere in the present work. In this section, I present a quick overview of such past formulations, though it is only a small sample of what could be an indefinitely long list.

One name for the idea is *the luck objection*.[132] If *E* is a matter of luck for an agent in the sense that whether *E* happens is not under the agent's control, then the agent has no free will with respect to *E*. Given this, one needs to argue that indeterminism causes such a lack of control. Notably, this problem also applies more broadly. Firstly, one can make the same objection against the compatibility of free will and determinism – in fact, it is more or less at the heart of incompatibilism in

[130]Doyle 2013, p. 237.

[131]Examples are too many to begin to list here, though see the next subsection. Examples of all these positions are found throughout this chapter and some others.

[132]See e.g. Moore 2022 for a summary, additional sources and further discussion.

general.[133] Secondly, it is just as intuitively plausible to say that the luck objection (from either indeterminism or determinism) applies to responsibility: if it is just luck whether *E* happens, one is not responsible for *E*. This ties back to the concept of moral luck (see 1.1.3).

Some authors discussed in this work have their own labels for their versions of the randomness argument. As mentioned in 1.1.2, Robert Kane refers to the problem of fitting free will together with indeterminism as "the intelligibility question".[134] Christian List speaks of "the challenge from indeterminism"[135] – his responses to it are discussed in 3.4.3 and 5.4.4. Peter van Inwagen speaks of "the *Mind* Argument" against the compatibility of free will and indeterminism, of which he has presented different versions[136], and later, he speaks of the "Indeterminism-Inability Principle" and "the Promising Argument" (of which the latter is an argument for the former)[137]. These all more or less amount to versions of the randomness argument, though I will not discuss them all in this work. Otherwise, van Inwagen's ideas are discussed in 3.4.9 and 4.1.

Perhaps the most thorough formulation of the randomness argument is found in 'Trick Slattery's book *Breaking the Free Will Illusion*[138], albeit with a popular and simplifying approach and hard incompatibilist motives. My argument takes a parallel path to his (without having been much influenced by it in the first place) and pays more attention to the kind of arguments talked about by philosophers. As a hard

[133]See e.g. 2.4.1, 4.1 in the present work.

[134]Kane 2002b, p. 18.

[135]List 2019, p. 109.

[136]Van Inwagen 1983, pp. 126–150, van Inwagen 2000.

[137]Van Inwagen, 2011.

[138]Slattery 2014. If you read this source, also note that Slattery usually uses the word "causality" when referring to determinism.

incompatibilist, Slattery also barely acknowledges the second main claim of the argument that I introduce below.

3.3.2 My formulation of the randomness argument: The two main claims

As I stated at the beginning of this chapter, the randomness argument as I present it can be seen as opposing two claims made by Robert Kane: that indeterminism in free will does not mean randomness or mystery, and that there are kind of freedom worth wanting that are incompatibilist. More precisely, its two main claims are the following:

- **The first main claim: Indeterminism in a choice always contradicts the agent's having control over the choice in a specific, important sense, at least in principle and to the extent that it has an effect at all.** Insofar as indeterminism applies to the choice, there is definitionally no reason why the exact choice that happens is the one that does happen. Hence, the agent's reason do not control the choice, and the agent themselves can only be the origin of the specific choice, if at all, in some way that still involves it being random. The specific sense of control involved in this argument is explained in 3.3.4. In the most simple terms: if a choice is undetermined between options *A* and *B*, then it is the case that the agent might choose *B* even if they have more reason to choose *A*. Now, it is true that there are various ways in which a choice can be indeterministic, and that only some choices might be undetermined, and the practical results can be quite different. Still, if there is indeterminism involved in the choice, insofar as it affects the choice between *A* and *B*, the possibility of choosing the worse option *B* always remains. It also applies just to the extent that indeterminism has a concrete effect; the only way to get rid of this problem is to minimise or eliminate the effects of indeterminism, which relates to the second main claim.
- **The second main claim: There is no reason to want free will to be indeterministic that does not reduce to the assumption that free will**

must be indeterministic (or equivalently, that free will cannot be deterministic). Since indeterminism leads to a lack of control in an important sense, or at best has no effect, there is no reason to desire it. Now, it is clearly the case that many people do desire indeterminism as a requirement of free will. They may also cite what seem like several different reasons for this. However, I will show that all such arguments are either mistaken or are different ways of expressing the basic incompatibilist premiss that free will requires indeterminism (equivalently: contradicts determinism). The second main claim is a negative claim and cannot be proven for good once and for all – there is always in principle the possibility that some new reason to desire indeterminism emerges. However, the first main claim gives much reason to suspect that it is true, and showing that existing arguments fail to provide any independent reason to desire indeterminism at least rules out many possible ways it could be like that and shows that those propounding those arguments are not basing them on any such reasons that they could articulate, and thus there is little reason to think such a reason would exist.[139]

3.3.3 Indeterminism as "randomness"

The problem with indeterministic free will and the intelligibility question will has been summarised by saying that undetermined events seem to "just happen", or be "arbitrary", "capricious", "random", "irrational", "uncontrolled", "inexplicable", or a matter of "luck" or "chance".[140] The word I prefer to use to get this idea across is here, of course, "random" – hence the *randomness argument*.

[139]Cf. Dennett: "We live our lives full of hope and striving, joy and regret, praise and blame. What about indeterminism would license any of this and what about determinism would subvert any of this? I have yet to see a persuasive response to this challenge." (Dennett, 2015, pp. xi-xii; also see p. 18.)

[140]Kane 2002, p. 18.

It is important to understand what I mean when I say that indeterminism is or implies "randomness". There are technical definitions for "randomness".[141] This has nothing to do with those. I am *not* arguing that from indeterminism follows something else, "randomness" defined in some particular way, and from that follow my conclusions. I am arguing that from indeterminism itself follow certain things, and those things are undesirable. Using the expression "randomness" is merely a way of *characterising* those consequences of indeterminism in an intuitive sense. In other words, "randomness" is only used for effect. If you would ask for its definition here, I would have to say that it is technically used to refer to the same thing as "indeterminism" – that is, indeterminism on the relevant level, which might vary but is likely to mean a level on which we describe choices as happening.

The use of "randomness" in this way is also a way of reminding everyone of just what indeterminism implies. Those defending indeterminism as a basis of free will are liable to forget or never see the implications. The word "randomness" recalls them to mind immediately.[142]

At the same time, we must be careful not to talk about different senses of randomness. This is discussed in section 3.4.1 below. Naturally, what I mean to talk about is based on the definition of (in)determinism I have been talking about.

3.3.4 Randomness destroys agency and control

Incompatibilists argue that determinism contradicts freedom because it implies that all of one's choices are determined by something outside of one's control (see section 2.4.1). Thus, libertarians want choice to be undetermined. However, this gives us

[141] See e.g. Eagle 2021.

[142] Trick Slattery (2014, pp. 74–75) argues that using the word "random" for indeterminism would be confusing because it could mean only apparent randomness. We should be fine because we already defined "indeterminism" first and use "randomness" in a different place, and in any case, "randomness" is strong enough to make libertarians usually want to deny it (see e.g. 3.4.1).

even less control.

Suppose you are choosing between alternatives, and your choice between those alternatives is undetermined. How much control do you have over the choice? If we see this as a libertarian free choice, we of course need to postulate that the undetermined choice that gets made is your choice. (See also 3.4.2 and 4.2.3 below.) Fine. So, in practical terms, how much are you able to rely on control of that choice?

If the choice were determined, you might be able to rely on it stemming entirely from appropriate desires, values, etc. that you have, hopefully through a good process of reasoning. However, we now suppose it is not. It is undetermined. This means that nothing is a necessary cause or reason for whichever choice is made. The final choice is something that happens for no reason at all.[143] There might be reasons why one out of the set of *A*, *B*, and *C* will be chosen, but not why *A* is chosen rather than *B* or *C*. Basically, within the limits that indeterminism applies, anything could happen. That is why I characterise the choice as random.

What if you want to choose *B* rather than *A*? All you can do, when the choice is indeterministic, is to hope for the best. You may be able to increase the likelihood of *B* rather than *A* being selected, if the indeterminism is something like statistical, but when it comes to the point when you make your actual choice, you can only wait and see what it is. You may, by postulation, also actively make the random choice. However, this is rather like saying you were the one who rolled the dice. It may be called your action, but you do not have control over it in any practical sense. Any desire for *B* you had, no matter how powerful... any principle you may seek to be

[143]Cf. Balaguer 2014, pp. 19–21. This discusses quantum physics, but it also brings up the point that, within a limited set of indeterministic options, whichever one happens happens for no reason. Ironically, pp. 23–24 make the analogous point about random choices, introducing the randomness argument, but instead of the wording of "no reason" and instead uses ambiguous expressions like "nothing caused it," "nothing made it happen," "it just happened" and "it was random." As we will soon see (3.4.1), this is problematic, as it ties the argument to potentially ambiguous or even false premisses. Of course, just saying "no reason" is ambiguous too and would not automatically fix the issue, but it does get at something important that I am expanding on right here.

committed to that implies choosing *B* over *A*... any reasoning you did that suggests *B* is the best option considering everything that is relevant in the unique specific situation... all of these can at most make *B* more likely. You cannot rely on them to enable you to trust that you will choose *B* – if you can, the choice ceases to be undetermined.

If choices, which we experience as being ours, are undetermined, we are in a peculiar position such that we feel that we are actively making a choice when no part of our minds other than a random choice generator has any control over that choice. Since this presumed random choice generator acts randomly, there is no practical difference between the results of our undetermined choice and a random choice made by a machine making random choices with the same odds. Yet, somehow, there is a metaphysical difference, and we are supposed to be responsible.

It has been argued that ultimate origination is more or less incoherent. To be the origin of something, you need to exist already (see 3.6). I am now making largely the same point as applied to randomness as freedom. Our desires, preferences, values, processes of deliberation and so on are an important part of who and what we are. Every time we say that (even) these things do not determine what we choose, we are saying that, in an important sense, we have no control over what we choose. Let us grant we have control in *some* sense because the random choices are our own random choices.[144] That gives us nothing helpful. The results are the same as if the randomness came from outside us. We still cannot guarantee that we choose the better option *B* no matter how much we might want to guarantee it. If nothing in the universe determines us, then we ourselves do not either. We may get to say we were metaphysically the cause, but we are still left with the problem that we may choose the less desired or objectively worse choice instead. The kind of practical control we are concerned with is not something that becomes unnecessary when you introduce some different type of control that does something else. You would still need to worry

[144]This will be properly discussed in 4.2.3.

that you may choose *A* for no reason.

All of this follows from indeterminism. Yet, understandably, libertarians have been disinclined to fully acknowledge it. Charles Hartshorne deserves credit for explicitly admitting that, in his view, freedom equates to "chance", which equates to "randomness".[145] Most libertarians do not embrace the label of randomness like this, nor are they keen on the notion of having to fear what they will randomly do next nor recognise that as the basis of responsibility, so the randomness argument should have some force against them.

It is important to note that this challenge concerns *higher-level* indeterminism (see 2.1, 11.3). To avoid this kind of randomness requires determinism under a higher-level psychological description. We need to be able to rely on our choices following from our motives and deliberation, not microphysical states. Thus, indeterministic microphysics that reliably coarse-grain into deterministic higher-level psychological phenomena (cf. 10.5 and 3.2.2) do not present a problem with randomness on the relevant level, at least provided that the lower-level indeterminism indeed never spills onto the higher level. A higher-level deterministic scenario with lower-level indeterminism also offers no help to the libertarian, since one still could not have done otherwise (it is not possible on the ultimate level). Conversely, lower-level determinism that become indeterminism on the relevant psychological higher level *does* have all the problems of randomness and loss of practical control. Thus, when Christian List[146] argues for a compatibilist interpretation of the ability to do otherwise based on a higher-level indeterminism, his model is close enough to libertarian to be affected by the randomness argument. To allow for a place for alternative possibilities in my compatibilist theory, I will need to be more subtle than this (see 5.4 and 9.4).

Though I refrain from defining (moral) responsibility yet, as I have pointed

[145]Hartshorne 1984, pp. 15–16, 18, 23.

[146]List 2019. See also 2.4.2 in the present work.

out, responsibility for one's actions or choices has been seen as intimately related to freedom and control (see 2.2.3). I will refer to such intuitions now and then when presenting the randomness argument. Assuming the typical intuitions, the randomness and loss of control implied by determinism also implies a loss of responsibility: how can you be responsible for a choice that was random?

I sum up the problem by saying that indeterminism threatens ***concrete control*** or ***practical control***. This means a person's control over their own choices or actions in the concrete sense that they follow reliably from the person's own motives.[147] Whatever other kinds of control may be supposed, the lack of concrete control means that there will be no reason why one choice happens rather than another, and thus also that there is no reason of the person's own why it does. Even more concretely: in a choice that is indeterministic between *A* and *B*, you might always choose *A*, even if *B* is much better or *A* turns out to be horrid on reflection.

3.3.5 Just a little randomness?

Have I been attacking a straw man? Surely nobody thinks that every choice of ours is completely random – and even though there is no true third option between determinism and indeterminism, I just pointed out above in 3.2.2 that indeterminism encompasses everything except complete determinism, and so there are plenty of options. Also, in 2.1 and 11.2, I take pains to define local determinism as separate from universal determinism, which also enables me to separate local and total indeterminism. If actual libertarian positions hold that, say, the indeterminism only holds when a person is making a choice between different desires, what is the point or force of arguments such as those above?

The point, in truth, is that the argument in the previous subsection was never

[147]It would be easy to jump from these words to equating concrete control with some form of the existing concept of *reasons-responsiveness*. However, I will take the long route of motivating and thereby illuminating this through a number of further arguments before spelling it out for the first time (outside this footnote in) 5.5.2.

only about complete indeterminism in choices. Even if the choice remains between *A*, *B* and *C*, all of which you also have reasons to choose, you cannot concretely control which one of them you choose – you can only wait and see. This also applies to other kinds of examples, which will be discussed below. The general point is that wherever in the system or process of decision-making indeterminism occurs, at *that* point, there is only randomness. Maybe it will be a small portion of the choice that is uncontrolled and random; fine, then, the rest is not uncontrolled and random, but that part is.[148] Insofar as control in the sense described above is needed for freedom, this addition of a little randomness makes the choice less free. Without the bald premise that indeterminism is required for freedom, it really does not add any freedom.

What is also notable about efforts to limit the scope of indeterminism is that they half concede the point of compatibilism – more, they half concede idea that determinism is *required* for free will. If you try to make your theory of indeterminist free will more and more like determinism in order to make it better describe free will as it should be, why do you think free will properly described would not be deterministic? I will return to this in section 4.3.1 below. Here we see that not only do we have contradictory intuitions, but the incompatibilist intuitions seem detached from what we want in practice.

All of this leads to a suggestion that I do not make seriously, and which I doubt anyone would be satisfied with. The ideal compromise between the bald demand of indeterminism and the reasoned demand of control would seem to be this: that free human actions should be indeterministic in principle but deterministic in practice by being "determined" only with an extreme high probability such that the possibility

[148]Cf. Balaguer 2014, pp. 25–28. Mind you, Balaguer here claims not only that a probabilistically "determined" choice is not only unfree insofar as it is random, but also unfree insofar as it is "partly" predetermined, and I do not agree with this latter. Not only does my definition trivially exclude something that is "partly determined" from being deterministic, but it is rare to see anyone, even an incompatibilist, consider "partial determination" as a threat to free will in the first place.

of doing otherwise exists in principle but will never be realised during the existence of the universe.[149] I say this, of course, to further highlight the strangeness of the demand for indeterminism.

The specific forms in which limited randomness has been or could be used in libertarian models will be considered in detail in 4.3.3 below.

3.3.6 How many choices are undetermined?

Libertarians can disagree a great deal about just which choices need to be undetermined for their incompatibilist requirements to be fulfilled, especially considering they also have the intelligibility question to deal with, thus having to avoid being *too* random. There are at least three general positions they can take on this.

- **Every free choice must be indeterministic.** The idea here may be, for example, that all the options need to be truly open for the agent to choose between them as a matter of necessity regardless of other details. This view is held, for example, by Peter van Inwagen[150] and Christian List[151] (see 3.4.9 and 5.4.4).
- **Choices with competing motivations must be indeterministic.** This means that choices where there is clearly only one thing that the agent would want to do are not indeterministic, but those where they have motivations to

[149]The problem with this concept is similar to the problem with the concept of *weak reasons-responsiveness*, and in fact there is still a similar problem in the concept of *moderate reasons-responsiveness* that is introduced to fix the problem in weak reasons-responsiveness. These concepts by John Martin Fischer and Mark Ravizza are discussed in 8.1.1 and 8.1.2 in the present work.

[150]Van Inwagen 1983, pp. 154–157.

[151]E.g. List & Rabinowicz, 2014. Note that for List, they only need to be undetermined on a higher level, not on the ultimate level. Still, the scope of which choices need to be undetermined is the same: every free choice.

choose more than one of the possible options are. Mark Balaguer calls these *torn choices*. He also adds the further requirement that only choices we make consciously count as these kinds of free choices, but these are still choices that happen several times per day for him.[152]

- **A few key choices must be indeterministic.** This appears at least in Robert Kane's idea of *self-forming actions*[153] (see 3.4.7). The idea there is to cut off the chain of deterministic causality going back beyond the agent at *some* point with a choice that is not determined by anything. These are torn choices of a sort, but major and rare. Most choices do not need to be indeterministic in this view, but as long as they follow deterministically from these indeterministic choices, they go back to the agent's choices about what to become like.

As I will argue below in more detail, and as I have already argued in more general terms above, none of these ideas escapes the randomness argument. Every choice needing to be indeterministic leads to the possibility of going against any interest of ours. Only torn choices being indeterministic still leads to the possibility of choosing the worse among your motives, except where it just makes no difference. Self-forming actions leave you directly affected by randomness only rarely, but those moments still lack concrete control, and their consequences reverberate throughout your life.

3.3.7 Opening the black boxes

I argued in 3.2.1 that there is no third option outside of determinism and indeterminism because anything that is not determinism is by definition indeterminism. It is important to notice that hiding a third option inside something

[152]Balaguer 2014.

[153]Kane 2002a, Kane 2011.

mysterious like a faculty of will[154] or a homunculus (see 10.7) within the brain is impossible. Philosophers, to say nothing of laymen, sometimes manage to hide this from themselves by wrapping it within confusing words, of which we can see examples below.

A process may be mysterious or unknown. It may be a sort of black box that cannot be opened. Nevertheless, it is always possible in principle – from perfect knowledge – to say that it is either deterministic or indeterministic. Given certain total circumstances at T_1, is there more than one possible state for it at T_2? If there is, it is indeterministic. If there is only one, it is deterministic. If we do not (or even cannot) know, we do not (or cannot) know which it is, but nevertheless, an answer exists by virtue of logic alone. The point here is not whether we do or can know. The point is that there is no logical option that would avoid both determinism and indeterminism, and thus conclusions correctly drawn about either determinism or indeterminism apply in each case, leaving no way to avoid both. We may not know which consequences apply, but we can know it is not the case that neither do.

3.4 Can the randomness argument be avoided?

The randomness argument as I presented it above is a blanket statement covering all forms of indeterminism. I gave an argument as to why it is able to do so. Nevertheless, in such cases, there is a danger of overgeneralising by not taking surprising alternatives into account. Libertarians might well claim that they have a theory or theories to offer that avoid the randomness argument.[155] Thus, in this section, I will go through as many such objections of different sorts as I can.[156]

[154]The idea of a faculty of will is explained, though without accusing it of being a black box, in section 1.2.

[155]The article Kane 2017 does exactly that – this source gives only the claim, not an argument.

[156]The reader should note that one argument for incompatibilism I will not consider separately in this section is appeal to intuition. Arguments of the form "We have an intuition that free will involves indeterminism (e.g. we experience that we could have

3.4.1 Equivocating on "chance", "indeterminism", "control"

Opponents of the randomness argument may claim that it equivocates on what it means by, among other things, randomness. One example of such a criticism is Laura Ekstrom's charge against van Inwagen's formulation, which from the quote provided by Ekstrom[157] can be summarised as follows:

1. "Libertarianism" implies that for acts to be free, they or their immediately causal antecedents must be undetermined.[158]
2. However, if acts or their antecedents are undetermined, how the agent acts is a matter of chance.
3. An agent can hardly have free will if how they act is a matter of chance.

"Chance" (which obviously is something like "randomness" in the randomness argument) can mean three things, according to Ekstrom. The first is when chance is seen as something like an agent, "a force with powers of its own"[159]. With this meaning of "chance", it might be that when chance determines what action is taken, this means the choice is made by another agent-like entity rather than the

done otherwise), and this is evidence that incompatibilism or libertarianism is true" *do not* counter the randomness argument even if they are successful. They aim to establish that we should define free will in a particular way, but they cannot establish that indeterminism can avoid the charge of loss of concrete control, or that there is an independent reason to want indeterminism from free will. That said, the following argument *is* possible: "We have intuitions that free will involves indeterminism. Therefore, given some additional premise, we can conclude that there is an independent reason to want indeterminism as a basis for freedom." Such an argument is discussed below, in 3.4.10. Other than that, intuitions and appealing to them in the context of defining free will are discussed elsewhere: 2.2, 3.1, and much of chapters 4 and 5. See also 6.3.1.

[157]Ekstrom 2011, p. 375.

[158]It might not be entirely correct for me to use "libertarianism" here, because that makes me not quite agree with this premise as I formulate it myself. (See 3.3.6 above.) One example of a theory that I consider as libertarian without fulfilling this requirement is Kane's (see 3.4.7 in the present work). Nothing much depends on this, however.

[159]*Ibid.*

agent whose choice it was supposed to be.[160] The second reading of "chance" is that a chance event has a probability of less than 1 of occurring. Ekstrom claims that an event may not have any numerical chance of occurring because it may be ungoverned by law; thus chance events in this sense are ones governed by probabilistic law.[161] The third reading of "chance" is that a chance event is "purposeless or not part of anyone's plan[.]"[162] This at least has some relation to loss of concrete control, but it is vague.

Based on these definitions, Ekstrom argues that van Inwagen's formulation of the randomness argument is not true with any one meaning of "chance". The first one, chance as an agent-like force, she admits to be so silly that it would be uncharitable to ascribe it to proponents of the randomness argument. In any case, van Inwagen's second premise would not be true under this reading. The second reading of "chance" as a probability of less than one is also untrue on the basis that some events may be indeterministic without having a probability. Even supposing that van Inwagen's premise 2 were true, Ekstrom claims this would not be harmful, because "it does not make those decisions purposeless, unguided, or haphazard." This is because "One may decide what to prefer for reasons that cause and justify, without necessitating, the decision outcome." The above pretty much also states Ekstrom's stance on the third definition of "chance" with respect to van Inwagen's argument. If "chance" means that chance events are purposeless or not part of anyone's plan, then the third premise that chance events are not free is true, but the

[160] *Ibid.*

[161] *Op.cit.*, p. 376. I am not sure there is such a coherent possibility as an event that may happen but has no numerical probability of happening, but fortunately, that question is not relevant in the present context. Also, I see no reason why there should be no possible alternative meaning of "chance" meaning happening indeterministically otherwise than probabilistically – how does it help with freedom to say your choice cannot even be given a probability of happening in a particular way? – but that does not matter here either.

[162] *Ibid.*

second premise that if choices are indeterministic they are matters of chance is not.[163]

How does this relate to the randomness argument as I am stating it? Firstly, the first main claim amounts to more or less saying that Ekstrom's definition 2 does imply her definition 3: being a chance event in the sense of (roughly) indeterminism does imply chance in the sense of not being due to the agent's own reasons. Secondly, her claim that we can choose for reasons that do not necessitate without the chance being too random is a major claim against the randomness argument in its own right. I discuss it separately in section 3.4.5.

We can certainly draw one lesson from this: the randomness argument need not equivocate between different meanings of "chance" or "randomness". It does, however, need to be clear about establishing a connection between them.

That all said, it should be noted that opponents of the randomness argument may effectively equivocate in the converse manner. Ekstrom is too explicit to be accused of equivocation as such, but her kinds of "chance" bring up the same question. Shortly put: faced with the charge that indeterminism means randomness and a lack of freedom, intended in the sense that indeterminism contradicts practical control and thus roughly the same as Ekstrom's sense 2, the libertarian may respond by arguing that indeterminism is compatible with freedom because it is compatible with some other form of control. (See 3.4.2 below, and 4.2.3 later.) What is needed is clarity with respect to what forms of control are being talked about at each moment – and what they imply. It has already been suggested that there are different senses of freedom or control that may be compatible with determinism and/or indeterminism. (See 2.4.2 above, and consider Kane's second claim at the beginning of this chapter.) However, the discussion of them so far has neglected insights gained by focusing on the randomness argument. This chapter and especially chapter 4 will show clearly just what the options are.

As another example, Mark Balaguer also presents some options for what

[163] *Op.cit.*: p. 376.

"randomness" could mean: unpredictability, uncausedness, there being no reason why this out of all the options happened, and not being caused by me (the chooser). He argues or at least states that only the last type would be incompatible with free will.[164] The randomness argument is based on showing why, specifically, the "no reason" interpretation is a problem.

3.4.2 Causal variants: Uncaused actions, indeterministic causation, agent causality

Since causality seems to play such an important part in questions of determinism and questions of agency, libertarians have come up with a number of different theories about how agency and causation relate to each other. Sometimes, these are explicitly offered as answers to the intelligibility question (see 1.1.2) or the randomness argument specifically.

This section discusses three kinds of theories that may be seen as clearly different alternatives in their proponents' view, but which all ultimately relate to the randomness argument in the same way. These three views are non-causal theories, indeterministic causation, and agent causality.

First, **non-causal theories**.[165] Though there is undoubtedly much to be said about these, in this context, the general idea is simple to state. What I mean by "non-causal theories" here are views in which an agent's acts are not caused by anything in the normal sense of causality, and are also indeterministic, but are nevertheless the agent's own acts.[166]

[164]Balaguer 2014, pp. 72–75.

[165]See e.g. Pink 2011.

[166]The converse of this would also be possible: one could present a view in which actions or choices are noncausal but deterministic. Then, you might say choices are not "causally determined", with an emphasis on "causally". This may be correct, but it is irrelevant here. A noncausal deterministic theory is a deterministic theory. My version of the randomness argument is not supposed to be against such a theory in the first place, and if the theory accepts that these noncausal deterministic choices are free, then it is a compatibilist theory and is not a libertarian theory.

Second, **indeterministic causation**. One thing the proponents of libertarianism like to point out against charges that indeterministic events are not in the agent's control is that *undetermined* does not have to mean *uncaused.* Thus, an undetermined event could be under the agent's control by being appropriately caused.[167]

Third, **agent causality**.[168] As an example of this view, a 2002 paper by Roderick Chisholm[169] recognises a form of the randomness argument and tries to give it an answer based on a different kind of causation. "Determinism" is incompatible with freedom because it does not allow doing otherwise, and "indeterminism" is even more obviously so, because if an act was not caused at all but just happened, no-one could be responsible for it. As an alternative to both, Chisholm suggests that free acts are not caused by any other events but are caused by the agent as a "substance" – a different but arguably intelligible form of causation. Thus, they are neither uncaused nor determined.[170]

Though there are other reasons to go into detail about these three views and their differences – particularly related to the requirement that free acts be caused by the agent and not something else – in the context of the randomness argument, all of them amount to the same thing. None of them is concerned with concrete control. All three postulate that the agent is the author of their act in some sense, but they do

---

It is possible that someone who was concerned with things being caused the wrong way, such as by physical causes rather than reasons, but not deterministically, would find this kind of a theory agreeable. The only place I have seen something like this come up is with Thomas Reid, discussed in 11.9.

[167]See e.g. Kane 2011, p. 394, Ekstrom 2011, pp. 373–374.

[168]Besides Chisholm, see e.g. O'Connor 2011.

[169]Chisholm 2002. I will here discuss the view expressed in this paper as an example, and I will not look into how Chisholm's views may have evolved since that time.

[170]Chisholm 2002, pp. 51–55.

not establish that concrete control is not lost due to indeterminism.[171] Noncausal theories cannot even claim that this is true in the sense that the agent (or their mental states, etc.) is the cause of the act. Indeterministic causality and agent causality can avoid the problem of uncausedness,[172] but not the randomness argument.

Consider Chisholm's claim that agent causality avoids both "indeterminism" and "determinism". I can easily concede that agent causation as Chisholm describes it is (if coherent) neither uncaused nor determined, but nothing in his argument counters my earlier claim that everything is either determined or indeterministic. Agent causation or not, in any such event, we can in principle open the black box and see whether it gives deterministic or indeterministic output.[173] In Chisholm's case, the answer is evidently indeterminism.[174] He denies determinism by stating that knowing all factors such as a person's desires could never be used to predict all of their choices other than with some less than certain probability[175]. He also asserts that this makes the person the ultimate origin of their choices, "a prime mover unmoved"[176]. Nevertheless, he fails to counter the randomness argument. His worry was that an "indeterministic" act would be "not caused at all, if it was fortuitous or

[171]Cf. Doyle 2013, p. 243: "But event *acausality* somewhere is a prerequisite for any kind of agent *causality* that is not *predetermined*."

[172]This assuming they are working models of what causation could be like in the first place, which I will not discuss here

[173]As noted under 11.9, Thomas Reid is a compatibilist by our definition while believing in agent causation.

[174]Derk Pereboom argues that agent causality *must* be "indeterministic", since it involves something being caused in a way that does not merely follow the laws of nature (2011, pp. 415–416). However, this is at least not true in the sense of "indeterminism" used in the present work for reasons discussed in 3.4.8; we could still ask whether the whole system comprised of both physical laws and the non-physical agent acts deterministically, or, if we like, whether just the non-physical agent does.

[175]Chisholm 2002, p. 56.

[176]*Op. cit.*, pp. 55–56.

capricious, happening so to speak out of the blue"[177]. The randomness argument shows that even if the act was caused in the sense meant by Chisholm, it could very well still be described by those other words. It would be, as I have been putting it, random. Locating the agent as a "substance" as the cause of the act would make the agent the origin of the act – in some rather strange sense[178] – but it would not restore concrete control. Things might happen with the agent as a cause, but there would still be no reason why one thing rather than another happened. You might still choose the option you did not want for no reason. The concept of something being a causal effect is never invoked in the (my) randomness argument, so restoring merely causality does nothing against it.

As with any other argument that seeks to answer the challenge of the randomness argument by speaking of some different kind of control, or avoiding some other kind of "randomness", it still matters that concrete control is denied. As per the definition of concrete control, if you do not have it, you will have practical problems where you will not be able to rely on your choices being ones that correspond to your interests. You will also be unable to avoid conceptual problems: sure, you have (ostensibly) established some form of "control" or "agency", but how much can we really speak of those things when there is a disconnect between your motives and your choice?

The control ascribed to the agent in agent causation and the other two causal variant theories does not offer a way to avoid the randomness argument, but agent causality especially does offer a way of understanding what libertarian free will

[177] *Op. cit.*, p. 51. Note how this wording suggests equivocating about words like "fortuitous" or capricious" again, as discussed in 3.4.1 above; in context, it effectively suggests that (only) if a choice is uncaused, it can be capricious etc., but if it is indeterministically caused, this problem can be resolved.

[178] Cf. Honderich 2002, p. 127, which asserts that the libertarians' "talk of a self or originator is talk of nothing clear." See also 4.2.3 in the present work.

could mean. I will make this point in 4.2.3.[179, 180]

3.4.3 Agential or higher-level indeterminism (List)

Christian List identifies himself as both a compatibilist in our sense and a "libertarian" in the sense of demanding alternative possibilities and believing they exist[181]. The trick is that his alternative possibilities are on a higher level, whereas the determinism that freedom is compatible with is on the ultimate level or at least a low level. The different options are available for the agent. List calls this ***agential indeterminism***.[182]

---

[179]Another minor point of view that plays around with causality to avoid the wrong kind of determination but focuses on consent as uncaused is one mentioned by Gary Watson. While not defending libertarianism as such, Watson suggests that the following view would be helpful for the libertarians' case:

> The historical dimension of the concept of responsibility results from the principle that one is not responsible for one's conduct if that is necessitated by causes for which one is not responsible. This leads to a problematic requirement that one be responsible for one's self only if one thinks of the self as an entity that causes one's (its) actions and willings. Libertarians can reject this view. What they must affirm is that we are responsible for what we consent to, that consent is not necessitated by causes internal or external to the agent, and that if it were, we could not properly hold the individual responsible for what he or she consents to. (Watson 1993, p. 143).

It is easy to spot at this point that adding supposed "consent" does nothing to change the lack of concrete control. Maybe we could say this idea leaves the agent lacking "concrete consent" in an analogous sense.

For one more causal variant – "simultaneous" causation as having to do with "essential cause" – and a refutal of its relevance, see Slattery 2014, chapter 8. As I have not come across just this idea elsewhere, I tentatively take it not to be a major strand of the discussion.

[180]Yet another view that could be considered here is that free actions are determined only by reasons, not causes. (See Young 1993, pp. 752–753.) The answer here is again simple: if the causes are or are based on something in the world, then to be truly determined by them is determinism by the definitions used, and if what is meant is not determinism, then the randomness argument applies as usual.

[181]List 2019b, p. 9.

[182]List 2019b, chapter 4. The idea of higher-level indeterminism atop physical determinism as making free will possible is also endorsed by Colin McGinn (McGinn 1994, pp. 87–88).

It is easy enough to see that this is not a way around the randomness argument, because no ultimate-level indeterminism is required in this model; so if the randomness argument does not apply to List's kind of "libertarianism", that does not mean it does not apply against what we have been calling "libertarianism". However, I wish to bring this up to point out that the randomness argument does (or at least can be extended to) apply even here... though it does not have to apply to every kind of higher-level indeterminism.

Suppose for a start that List means that, on a higher mental level of description that contains all the pertinent details, considering all the factors exactly as they are in a given situation, a person's choice is still undetermined by everything in that description. If you look at what has been said about the randomness argument so far (see also 3.3.3), you will see that this is just what has been meant by "randomness". Given the situation and your motives and so on, you could do otherwise for no reason – thus, you lose concrete control. It is this that matters, not what happens on the ultimate level, though what happens on the ultimate level may of course have clear implications for what happens on a higher level. Thus, if this is what List means – and it seems that he does[183] – he has managed to come so close to libertarianism in his compatibilism that he even runs afoul of the randomness argument.

At the same time, List is unlikely to avoid the problems seen by most incompatibilists. They could still complain that in his model, regardless of the higher-level indeterminism, it is fundamentally true that it is not possible for us to do otherwise, our choices are determined by events long ago, and they are only weakly emergent on impersonal physical events.

The above notwithstanding, it will turn out that indeterminism on a certain higher level *does* fit together with the compatibilist picture of freedom I am sketching – and is even a requirement for freedom under it. (See 5.4.) As part of this argument, I am even going to present a model developing Christian List's ideas further to see

[183]List 2019b, pp. 81–86.

how what is good about them can be gained without the problems – see 5.4.4. The essential point is that in this latter case, the indeterministic higher level is one that generalises over multiple situations – not indeterminism in the individual situation as described in terms of all the agent's mental states and intentions. So while I am going to say later that there is a kind of indeterminism on a higher level that is worth wanting for freedom, this indeterminism is of a kind built atop determinism on a lower level, so while it might work for a "libertarian compatibilist" in the lines of List, it does not help libertarians as I have defined the term here.

3.4.4 What about determinism on the higher level only?

Incompatibilists are generally concerned with determinism on the ultimate level (see 1.3 and 10.6). This might lead to the thought that if they will not accept List's solution above of determinism on the lower level and indeterminism on the higher level, maybe they could accept indeterminism on the ultimate level coupled with determinism on the higher level.

Determinism on the higher level *could* preserve concrete control, but it does not seem that it should be an option acceptable to incompatibilists. After all, indeterminism is supposed to preserve the ability to do otherwise (see 2.4.2). If our actions are determined on the higher level where they are described as actions, then we cannot do otherwise. Our choices cannot be different than they are; this is unaffected by the fact that the details that realise the same choices on a lower level of description can be different. There are other senses of being able to do otherwise than strict possibility, but as we have seen in the section just mentioned, those are part of a compatibilist approach, not libertarian.

That said, I can point to at least one example of someone expressing a view that determinism on the higher level only might not threaten freedom. In his *An Essay on Free Will*, Peter van Inwagen writes as follows:

> It may well be that, for all that is said in this book, human behaviour is wholly predictable on the basis of laws that are about the voluntary behaviour of rational

> agents. Moreover, I see no reason to think that such predictability would be incompatible with free will.[184]

The context for this surprising concession is as follows: van Inwagen argues that it is laws of nature* on a physical level, not on a level of the intentional actions or rational agents, whose determinism threatens human freedom.[185] Additionally, van Inwagen contends that this kind of determinism does not threaten the *ability* to do things in the way in which one will never do it.[186] Thus, based in what he says here, his conception of ability is a compatibilist one at this point, while his incompatibilism is limited to the non-psychological lower levels, presumably in particular the ultimate level. However, van Inwagen is not consistent in this position, since he also asserts in the same book that agents must have multiple options accessible to them in an indeterminist sense[187], which is a statement of higher-level incompatibilism. Van Inwagen's view is discussed in more detail in 4.1; also see 3.4.9.

Whether such a view is actual or not, against a version of incompatibilism where higher-level determinism is acceptable, the part of the randomness argument that relies on loss of concrete control loses at least some of its force, since lower-level indeterminism does not mean a loss of control if it is accompanied by appropriate higher-level determinism. However, this idea of free will being compatible with determinism only on the higher level is still vulnerable to the objection that there is nothing independently desirable about indeterminism (even on

---

[184]Van Inwagen 1983, p. 64.

[185]In the vocabulary he uses, these higher-level laws are not defined as "laws of nature" (*ibid.*, 64) and therefore cannot constitute "determinism" (*ibid.*, 65). As usual, I use the terms in the sense I have defined them myself, and my point stands that way.

[186]*Ibid.*, pp. 64–65.

[187]*Ibid.*, pp. 86–91.

the lower level only) for freedom. Whether there is such an independent reason will have to be established through other arguments.

Any other arguments about determinism or indeterminism only on a higher or lower level face the same objections if they try to get past the randomness argument. Much can be done with different levels of description, but any level is going to be either deterministic or indeterministic. In particular, anything emerging on a higher level has to be one of these, and if it is indeterministic, then the randomness argument applies. For example, John R. Searle suggests in "Free Will as a Problem in Neurobiology"[188] that randomness on the quantum level could emerge as rationality on the level of the brain as a whole. He assumes free will worth talking about is indeterministic[189], and that decisions cannot be physically determined by the state of the brain before they happen[190]. However, he also thinks quantum indeterminacy is just "randomness" that cannot be freedom. Instead, he thinks the properties of the system as a whole on the higher level built atop this random level could have non-random indeterminism, though he does not particularly explain what this would mean.[191] My answer to this is predictable: certainly there could be something less random than quantum events are, but if it is still indeterminism, the randomness argument does apply.

3.4.5 Choosing among your own reasons

One natural way of thinking about how we choose in a libertarian sense is that, in contrast with acting completely randomly, we have various reasons to act in various ways, and we choose which of these to follow. Of course, a compatibilist might say

[188] An essay based on a presentation, published as the middle (second) chapter of Searle 2007.

[189] Searle 2007, p. 47.

[190] *Ibid.*, p. 58.

[191] *Ibid.*, pp. 74–76.

this too, but for the libertarian, that choice must be undetermined.

An example of this kind of conception of choice is found in Thomas Pink's article on non-causal theories of agency.[192] Pink distinguishes the case in which causal powers may be exercised in more than one way, perhaps probabilistically, from the case in which an agent determines what will happen by exercising freedom. In the first case, what happens is pure chance and random, but in the second, the agent determines what choice is made.[193] Taking this claim at face value, we have a way of escaping the randomness argument.

Do we, though? Can we instead claim that the randomness argument applies because the choice made by the agent in this scenario is still either deterministic or not? The answer is yes, we can say that. It is perfectly intelligible to ask whether the choice made by the agent could have happened otherwise in the strictest sense or not. If it could not have, then it was deterministic. If it could not, then it was indeterministic. If it was indeterministic, it was still "random" in the sense of contradicting concrete control. If we follow Pink's claims to their logical conclusion, we are to take it that there is an indeterministic event that is called "not random" (etc.) because it is up to us in some hard to define way, even though it is indeterministic and random in the context of concrete control and the randomness argument. We will look at what way this action could be ours in 4.2.3, but so far, it suffices to say this in no way avoids the randomness argument. This scenario still has us at the mercy of a random event that is called our choosing. There is no reason for our making the one choice rather than another. Quite possibly, Pink does not realise this, and instead imagines a scenario in which we choose and the question of determinism is not addressed by asking whether the choice is deterministic or not, but just because this may be psychologically possible to imagine does not mean that

[192]Pink 2011. For more on noncausal theories (or at least how they are irrelevant for the randomness argument), see 3.4.2 in the present work.

[193]*Ibid.*, 364.

the question cannot be asked. Pink presents no meaningful difference in the randomness of causal natural law and the randomness of what is postulated to be an occasion of choosing. Instead, he merely uses a different vocabulary in describing them.

It is true that if the indeterministic event happens in selecting between our own reasons, then at least whatever happens is caused by one of our own reasons; this leaves more concrete control than if we could randomly do something we would have no reason or motivation for at all. However, why would it be less free to choose the best among our reasons via some reliable process that might not randomly choose otherwise?[194] Concrete control is still lost at this point.

In summary, postulating an agent choosing between their own reasons does not avoid the randomness argument because the choice is still either deterministic or indeterministic, and the same things follow from indeterminism as in other cases.

Section 3.4.7 below discusses Kane's theory about self-forming actions[195], which is also a form of "choosing among your own reasons" but contains further details and may involve and admission that the choice is random.[196]

---

[194]This point becomes relevant in section 3.4.7, and then again in 3.5.

[195]Kane 2002a, Kane 2011.

[196]Balaguer (2014) presents a theory of libertarian free will where he specifically states that only some choices need to be free in an incompatibilist sense: those where the chooser is torn between different motives. (This is also mentioned in 3.3.6 in the present work.) See also Balaguer 2014, pp. 81–84, which puts an odd spin on this. It postulates an example case where neuroscientists saw what neural events happened in your head when you made a decision and they also saw that nothing caused which decision you made. It then postulates someone saying that this means the event was random and not your choice, and answers this by saying the neural event *was* your choice. The odd spin comes from Balaguer next arguing that because, in the materialistic view, the neural event is your conscious choice, it is not right to say that there was some randomness uncontrolled by you. Thus, it is as if he is suddenly answering an implied demand for a homunculus explanation (see 10.7 in the present work), but treating that as an answer to the randomness argument. This makes a kind of sense due to the wording of the particular example, but only that way. I would agree that it seems choices are something like neural events – though I am not committed to this in this work, see 2.1.6 – but I would be concerned if the nature of these neural events was to be concretely uncontrolled.

3.4.6 Two-stage models

There is a type of libertarian theory that can be shortly stated as follows: the choice that the agent makes between different options is not indeterministic, but the whole process of making the choice is indeterministic because the selection of options that the choice is made between is indeterministic. Robert O. Doyle calls this the ***two-stage model*** and says that such theories have been formulated though not necessarily endorsed by at least the following authors: "William James, Henri Poincaré, Jacques Hadamard, Arthur Holly Compton, Karl Popper, Daniel Dennett, Henry Margenau, Robert Kane, David Sedley and Anthony Long, Roger Penrose, David Layzer, Julia Annas, Alfred Mele, John Martin Fischer, Stephen Kosslyn, Storrs McCall and E. J. Lowe, John Searle, Uwe Meixner, and Martin Heisenberg."[197] He also proposes such a model himself, which he also claims to be scientifically based.[198]

In Doyle's words, the two-stage solution solves the problems posed to freedom by both determinism and indeterminism as follows:

> Limiting indeterminism to the first stage prevents it from making our decisions themselves *random*, which would threaten our responsibility. The "determinism adequate" of the second stage defeats the problem of *predeterminism* from the Big Bang that threatens our freedom.[199]

In other words, this solution is supposed to address the randomness argument by making the decision between options be not random, not indeterministic, while

---

[197]Doyle, n.d. Though this source is not a published article or book, it presents a good overview of the topic as well as introducing a good term for it (that is, *two-stage models*). Meanwhile, Doyle does have publications about the same and related topics, such as Doyle 2013.

[198]Doyle 2013.

[199]Doyle 2013, p. 237.

also avoiding the threat to freedom posed by predeterminism by the distant past.

If we are not convinced that this is enough to defeat the randomness argument, how might we respond? Consider Mele's version: What we decide is influenced by what comes to our minds while we are considering the decision, and even if it is deterministic which things do come to mind, we still do not have control over that, so we lose nothing in terms of control if it is indeterministic.[200]

While this rightly points to how full ultimate origination is compatible with neither determinism nor indeterminism (see 3.6), the indeterminism that is supposed to be a requirement for freedom ends up being the possibility of some things not coming to mind. Thus, we are not determined to choose the option that our reasons would most dictate because we might not think of the relevant factors. How is this a picture of greater freedom than if we were determined to think about it?

The above question concerns both claims of the randomness argument. Firstly: Yes, there is a lack of concrete control in the situation in which the earlier stage of the process if random. You would have more concrete control if some potential options you could choose got a chance to be considered, instead of being randomly ignored. If the randomness argument states that inserting indeterminism anywhere into the process also causes there to be a corresponding lack of control, it does not help to add control in another part of the process. Certainly it may make the whole process less random, and give more concrete control, but it is still true that the part that is indeterministic represents a lapse in concrete control. It might be that the randomness plays such a small part that it does not matter, but that is a whole other argument – see 4.3.1 in the present work.

As for the second main claim, the two-stage solution does not seem to add anything independently desirable by adding indeterminism. It seeks to solve the problem of predeterminism, a problem that is just a more specific version of the claim that indeterminism at some point is needed for freedom, but adding

[200]Mele 2002, pp. 543–5.

indeterminism in the earlier stage does not create a scenario in which we could see that indeterminism is somehow helping in any other way than by fulfilling the requirement for indeterminism. Rather, the lack of concrete control shows that there is a way in which the indeterminism hurts.

However, there will be more to say about this question. Using randomness for generating a selection of ideas to choose from might be concretely useful as well, in ways not yet touched. I will return to this idea below in 3.5.

3.4.7 Kane and self-forming actions

> If it isn't worth doing, it isn't worth doing well.
> –Donald Hebb according to Daniel C. Dennett[201]

The two-stage solution locates the indeterminism in a specific place in the process of decision-making while postulating determinism elsewhere in the process. Obviously, it is possible to do this in other ways – locating the indeterminism elsewhere in the process. One such way, which also seeks to counter predeterminism, is proposed by Robert Kane.[202]

Kane's libertarianism does not demand that every choice be undetermined. Rather, he wants to find a way in which we can be the ultimate origins of our own choices. Obviously this is difficult because he wants our choices to both flow from our own reasons and, at the same time, not be predetermined by anything that came before us – but it seems our reasons must come from something that was before.

Kane's solution is the following. We can be responsible and ultimate origins of our own choices if we make some choices that are not determined and that help

[201]Dennett 1991, p. 460. Dennett is not using the quote to comment on the same thing as I am here, and I do not know the original context in which Hebb produced it.

[202]Kane 2002a, Kane 2011.

shape our character so that our future choices are in part determined by them. At the same time, these choices have to follow from our own reasons. To fulfil these criteria, these character-determining choices would happen in cases when we are trying to make a difficult decision with strong motives for both of two contradictory choices. Kane speculates that, in such a case, our brains might go into an indeterministic state, such that either of the two motives might become realised. The choice made in such a state would then follow from our own motive (whichever one it was), and the choosing would not just be some random event, it would be our act, because it would happen through our own effort.[203]

Such an ostensibly undetermined choice is exemplified in the short story "The Lady, or the Tiger?" by Frank R. Stockton[204], though this example is not used by Kane. The story is set in a kingdom ruled by a "semi-barbarian" king. Among the king's whims is to subject accused (male) criminals to a sadistically entertaining trial by luck, with the excuse that God will see to it that justice happens based on whether they are innocent. The accused is faced with two doors and has to choose one. Randomly assigned, behind one door is a tiger that will kill the accused, and behind the other, a beautiful woman whom the accused man will marry if he chooses that door. When the king's daughter is caught having an affair with an unsuitable man, the king decides that putting the man on such a trial is a good way to get rid of him. However, the king's daughter has found out which door leads to the lady and which the tiger. Based on this knowledge, she gives a signal to her lover from the audience, and he chooses the door she indicates. However, the story shows the internal agony of the king's daughter as she finds equally intolerable either the image of her lover being with another or being killed by the tiger. Thus, the story ends with an impossible to answer question: which was behind the door the man opened, the lady or the tiger?

[203]Kane 2011, p. 387.

[204]Stockton 1882.

Of course, the indecisive character in the above example is peculiarly amoral and only considers her own emotional reactions in spite of another person's interests being heavily involved. A more typical example by Kane might involve actual moral considerations fighting with the temptation to follow one's self-interests, as in the case of the ambitious woman who could either stop to help or hurry on to make it to a meeting on time[205].

Kane seems to[206] eventually admit to certain randomness in his self-forming actions when he states that each of them is a value experiment in a sense he has introduced elsewhere.[207] The idea of a value experiment is that a person is effectively saying

> "Let's try this. It is not required by my past, but it is consistent with my past and is one branching pathway my life can now meaningfully take. Whether it is the right choice, only time will tell. Meanwhile, I am willing to take responsibility for it one way or the other."[208]

This is a clever move in that it dreams up a way in which a person could be intuitively[209] responsible for a random choice: if they made a choice to choose a

[205]Kane 2011, p. 387.

[206]I find it hard to say what exactly he is explicitly admitting here, because he is verbally admitting that there is some truth to the charge that self-forming actions are "arbitrary", and seems to present that as a real admission of a weakness, but he next goes on to connect "arbitrariness" with its roots in Latin that have to do with free choice, which seems like much less of an admission. (Kane 2002a, pp. 236–237. Cf. Kane 2001, p. 401.)

[207]Kane 2011, p. 236.

[208]Kane 2002a, p. 236; quotation marks in the original. See also Kane 2011, p. 401.

[209]I do not disagree that they would be really responsible, too. I just need to say "intuitively" because I need some standard I can appeal to when I have not yet introduced any (see 2.2.3).

course of action randomly. However, this hardly describes the situation Kane originally described. That situation was "An agent wills both *A* and *B* and one or the other is randomly the one that happens." The value experiment for which a person would be intuitively responsible would be "The agent freely chooses (whatever that means) *V*, where *V* is 'Randomly select either *A* or *B* and perform it.'" Kane has been talking about a situation in which the person wills *A* and wills *B*; there has been no *V,* nor willing of *V*. We cannot (intuitively) be held responsible for choosing to randomly choose if we have not chosen to randomly choose. There is no evidence (nor does Kane suggest there is until this point in the argument) that we are in the habit of thinking that we should choose a motive randomly. Indeed, that is the very opposite of the struggling he describes, since choosing to randomly choose involves indifference and struggling involves strong motives and values. Hence, the line of argument that we are responsible for random choices as value experiments we choose fails.

Kane's contention that these choices are ours because we make them through our own effort is not meaningful. It is only true in much the same sense as if we had to expend a great deal of *effort* to move a heavy lever without knowing what it will do.[210] If we really make two contradictory efforts at the same time and one randomly happens[211], why is the one that actually happens more our own than the other one we were also trying to bring about? Perhaps it is our fate or our circumstance, similarly to something imposed from the outside that we may have to accept, but why is it our

[210]Cf. Honderich 2002, p. 52, speaking of Kane's theory: "But if the various verbs and locutions [such as "efforts", "struggles", "willings" etc.] are deprived of a standard causal content, which they must be, and given only some content having to do with probabilities, the choices and decisions remain unexplained. For all that has been said, any one of them might never have happened." The same applies even if we add more words by which we can describe them, such as "attempts", "endeavourings" and so on. I am not sure Honderich's way of putting it is quite right, though. Perhaps, in Kane's scenario, we do have multiple endeavours, or whatever, going on at the same time. The real problem is that none of them is more our own than the other, and then we come back to the point about probabilities Honderich mentions.

[211]Kane 2002a, p. 231.

own product or responsibility?

The same problem comes up if you express the argument by saying that we are responsible for whichever choice we make because it follows from our reasons either way. If we are responsible for the choice because we had our reasons to do it, then we should be equally responsible for the other possible choice we did not make because we also had reasons to make that choice. This is further highlighted by pointing out that often the responsibility could be phrased as being about choosing *A rather than B*. If we are responsible for the choice we actually made rather than the other one because we actually made it, and we actually made it due to randomness, then we are responsible for what randomly happened. Unless Kane were to say we are responsible also for the choice we did not make, his model does nothing to help. We are still at the mercy of luck. One might say we are not responsible for choices we could have made but did not, but that is only a matter of formulation, for it would not be plausible to say that we would be responsible for making one choice rather than another if it was random which one we did. It does not really help to say that this random chance partly determines our future character and choices as well; that only leaves us more dramatically at the mercy of these few chance choices.[212]

[212]Compare this with my personal response to "The Lady, or the Tiger?" Obviously my response is that the answer to the question is objectively unknowable, but subjectively, I would answer "The tiger, or might as well assume so anyway." The point is that if there is an equal chance that you cannot trust someone in an important matter as that you can, and especially if you consider being in a relationship with such a person, it does not matter much if they coincidentally happen to choose what is right. (I take it that the tiger was the morally wrong option for the character to choose.) Thus, you might say, if it was not the tiger this time, it would be some other time. I do not necessarily deny the impulse to hold the character responsible for whichever option she actually chose, but there is a sense in which happening to choose right in such a case is much less meaningful than a choice that is determined to be the right one by better values. In particular, if she made the morally wrong choice and was held responsible for that, a great part of that responsibility could be seen to be due to the fact that she did not have the initial decency that would have made it determined that she not do it (cf. Hobart 1934) – much more than the fact that she happened to get this result rather than the other from the indeterministic roll of the dice.

The basic answer to Kane's theory as a counterargument for the randomness argument is the usual: at the point where indeterminism is inserted into the model, at that point, there is no concrete control, and thus, it also seems strange to postulate responsibility. There may be at other points in this hypothetical model of the decision-making process where there is practical control, but as soon as indeterminism steps in, that step becomes practically uncontrolled. Again, indeterminism adds nothing but randomness.

Kane is one of the most deft players I have encountered in this game of trying to embrace indeterminism but avoid its consequences (see 4.3.1 below), but I think that he is trying to do something that cannot be and should not be done. What real insights he has in my opinion (such as the human significance of some kind of self-forming actions) are unrelated to indeterminism. Kane's answer involving self-forming actions is little different in this respect from appealing to limited incompatibilism plus agent causality.[213]

3.4.8 Dualism and "game rule" indeterminism

Since one motivation for objecting to the compatibility between determinism and free will is to object to something like external events or laws of nature as being the causes of actions, rather than the agent's own intentional actions (see 2.4.1), we could also raise an argument to the effect that having something like a dualistic system with the mind/soul separate from the laws of nature would help with the problem of free will and (in)determinism. Maybe nature is deterministic but the soul is not?

Countering the use of dualism as an objection to the randomness argument is a rather simple matter, especially in the light of what was said in 3.4.5, so I will combine the answer in this section with a particular form of dualistic model that also involves a novel variety of indeterminism.

[213]Incidentally, Kane states that theories of the sort his is also involve their own kind of "agent causation", yet are different from actual agent-causal theories (Kane 2002a, p. 239).

Nicholas Denyer discusses a dualist libertarian alternative that might sound as though it could offer a new option among the variants of determinism and indeterminism:

> It is quite compatible with this refutation of physical determinism that the laws of physics are all deterministic in the sense whereby deterministic laws are contrasted with statistical ones. The laws of chess are in something like this sense deterministic rather than statistical. They absolutely rule out certain ways of moving the pieces, and not one of them takes a statistical form, requiring in some circumstances that some move be made only a certain proportion of the time. The laws of chess however do not form a deterministic set in the sense that given any one configuration of pieces on the board they allow only one configuration to follow; for in all save rare circumstances, they leave several alternatives open. The laws of physics are then, if deterministic, deterministic only in the sense that that the laws of chess are.[214]

In our terms, the rules of chess are indeterministic because more than one future alternative is possible in most cases. However, it is true that they are not statistical, and also that they are rules of a sort.[215]

A similar point is made by David Hodgson[216]. Hodgson considers three hypothetical universes. One of them is found to run by the rules of chess, the second by the mathematical "Game of Life", and the third by something he calls Superlife. In all cases, the universes are being observed by hypothetical scientists who infer the presence of different features in the universe's workings by looking at the laws it

[214]Denyer 1981, p. 98.

[215]Anthony Kenny (1973, p. 103) also points out the possibility of "chess-like" rules, claiming that philosophers seem to believe in determinism because they confuse exceptionless rules (as of chess) that are always obeyed by the system with "complete", e.g. deterministic ones. That said, Kenny is not defending libertarianism or incompatibilism.

[216]Hodgson 2002.

seems to follow.

The chess universe is found to work according to the rules of chess. What happens there is never contrary to those rules. However, it is also found to be "played" fairly intelligently. Out of the moves that are possible in a given situation, the ones that happen make sense if they were made by a chess player of some sort. This points to a dualistic model where minds outside the materially visible universe are playing chess with the intent to win within the possibility of the rules.[217]

In the second universe, the universe is found to work according to the rules of the Game of Life, a mathematical cell automaton developed by John Conway. This "game" consists of an infinite, two-dimensional grid of squares (cells), each of which can be in one of two states, "alive" or "dead". The system is governed by simple rules, according to which each cell's state in the following round is determined by the amount of living cells surrounding it in the current round. The system is deterministic but can display surprising large-scale patterns and sensitive dependence on initial conditions, making it a handy example of emergence for those who accept it as such[218]. For Hodgson, this universe is an example of what ours could be like if determinism is true, or even if quantum indeterminism that largely cancels out at large scales were added out to the mix.[219]

The third universe is the Superlife universe, which could in fact be our own. In it, the scientists observe that the behaviour suggests partly deterministic and partly statistical rules. The universe also contains systems of elementary particles that appear to behave like purposive agents, and it cannot be conclusively shown whether their behaviour is deterministic and whether it is indeed purposive. Two hypotheses would be formed about this universe: either all of the behaviour of the universe is governed by deterministic and statistical rules and all variation from determinism

[217]Hodgson 2002, pp. 249–251.

[218]E.g. Cohen & Stewart 1994, pp. 214–217.

[219]Hodgson 2002, pp. 253–255.

(due to to the statistical nature of some of the rules) is "random"; or the apparent agents within the universe are able to make genuine choices within the leeway given by the rules. The fact that the agents seem to experience choice, as in our world, could weigh in favour of that interpretation.[220]

From the point of view of the randomness argument, the answer to dualistic models in general is simple. If there is a dualism, with the material being one of two kinds of substances, fine – but does that other substance behave deterministically or not?[221] It may be psychologically easy to skip that question by thinking of making a choice instead of thinking about the question, but that is just another homunculus (see 10.7, in the present work). There is an answer to the question in principle, even if the second substance is a black box that we cannot see inside. With each choice made by the second substance, we can ask whether it could have happened in more than one way or not. If it could have happened in more than one way, then if this has any effect at all, it is to lessen concrete control.

In other words, dualism changes nothing[222], and neither does the idea of indeterministic, non-statistical rules such as in chess. What is more interesting about the idea of the chess universes is to ask whether the level of choosing between the "different moves" might not be deterministic but still help with freedom somehow. (See 5.4.4.)

I defined (in)determinism based on possibility within laws of nature*, already making sure they do not have to be exactly what are normally called "laws of nature". Someone proposing a dualistic solution might contend that whatever laws might

[220]Hodgson 2002, pp. 253–255.

[221]This point is also acknowledged by Peter van Inwagen (2002, pp. 191–192), Colin McGinn (1994, p. 84) and Manuel Vargas (2013, p. 41).

[222]This point is also argued, more thoroughly, in Slattery 2014, chapter 29, and briefly in Balaguer 2014, pp. 41–42; the reader will know at this point that they can replace "caused" with "determined" in the wording of these sources to make the argument sound.

govern the non-physical realm, or whatever the other half of dualism is, are not laws of nature but something different. To this, my reply is: does any of that really matter? All that matters is that we talk about non-logical laws that govern how some part of the world works. The logic of the randomness argument does not depend on how the laws that count within are termed, as laws of nature or otherwise.[223, 224]

Finally, remember that if what you are really concerned with is not so much this simple definition of determinism as the idea that human behaviour would be governed by physical or natural law, my argument does not assume that it is, as explained in 2.1.6 (and see also 11.9).

What about the "game-rule" models where (dualistic or not) the game or game-like thing is governed by rules that are neither probabilistic[225] nor deterministic? Well, they are still indeterministic or deterministic depending on whether the agent "playing the game" does so deterministically or not. Their choices are constrained by the rules of the game, but other than that, there is nothing going on that is different from any other choice in this respect. If the rules allow multiple different moves in a given scenario, then you have to choose between those moves, and either it is only possible that you make one specific move (determinism), or there are multiple choices that are all possible until you have made the choice (indeterminism). Everything that has been said so far about the problems of that choice being indeterministic apply here. There being partially limiting rules really does not do any work other than making it potentially sound as though relevant alternative different from standard deterministic and indeterministic options has been

[223]Cf. Slattery 2014, p. 322, endnote 8.

[224]I bring this up because so much of the discussion so far has consisted of finding loopholes by latching onto avoidable problems such as this, as shown in the rest of the present section (3.4).

[225]It is not clear to me whether you could really avoid a probabilistic model if you were to explain in detail what it would mean for non-deterministic moves in a game to be like something that an agent could have made.

found. Something interesting may come of it (again, cf. 5.4.4), but not an escape from the randomness argument or the dichotomy between determinism and indeterminism.

In the next subsection, we look further at what it would mean to deterministically or indeterministically deliberate between options, whether in such a "game" or otherwise.

3.4.9 Belief in determinism excludes deliberation?

Peter van Inwagen argues that consistent belief in determinism makes it impossible to deliberate about what to do. If this were true, it would be a counterargument to the second main claim, as deliberation certainly seems independently desirable. Roughly same argument has also been advanced (at least) Alan Donagan[226] and Christian List[227], but I follow van Inwagen's version here insofar as there is any difference. I will return to List's take in section 5.4.4.

Simply put, van Inwagen's argument is that in order to intend to do something, one must believe that they are capable of doing it, and so, to deliberate between actions, one must believe that one is capable of doing each of the possible actions. If one believes in determinism, then, given the arguments and definitions van Inwagen gives elsewhere in the book about determinism and ability to do otherwise[228] – or just given PAP on the ultimate level – then one knows one cannot do more than one thing, whichever one it will be.[229]

This is not quite saying that determinism itself makes deliberation impossible, but since consistent and true beliefs are better to have than inconsistent and false

[226]Donagan 1987, p. 170.

[227]E.g. List & Rabinowicz 2014.

[228]See especially chapter III in van Inwagen 1983.

[229]Van Inwagen 1983, pp. 154–157. See also van Inwagen 2002, pp. 193–194.

ones, indeterminism would be desirable at least to make deliberation possible while maintaining true and consistent beliefs about determinism and indeterminism.

Is it true that deliberation requires belief in indeterminism? I will say here that I think that in some sense it requires being able to think of the options as potentially (at least counterfactually) *open*; this is discussed again in 5.2.1. Here, though, I will take a different approach. What do we want to do when we deliberate? Probably not to be random, or even spontaneous; in deliberating, we are trying to find a good option to choose. Given that, which of these would we rather have?

1. We will deliberate, come to a good solution, and execute that.
2. We will deliberate, and then do one thing or another, which is undetermined by considerations of what is a good option.

The first option allows the process to be deterministic – even requires determinism if we think in terms of the single best option being selected – and the second requires indeterminism. The second one is not desirable, nor the point of deliberation. If we want to be random, why want to deliberate about it? And: if we deliberate about what to do, why would we want not to be reliably guided to the best option?[230]

If we see deliberation as a process of choosing between options, where we are able to choose the best one (to some approximation), it does not require indeterminism; it does not even require "up-to-usness" or "freedom" *in the sense that van Inwagen defines them*, and those turn out to be a hindrance instead, logically leading only to reason to fear doing the wrong thing for no reason. We can believe we are capable of going through a process of deciding that leads to a single outcome, and to the extent that this is the best outcome, we have no reason to desire the process to be different. In fact, van Inwagen himself has later come to a similar but weaker conclusion – not embracing the determinist conception but recognising the difficulty in the indeterminist conception partly[231].

[230]Cf. Balaguer 2014, p. 64.

[231]Van Inwagen 2000.

Van Inwagen's ideas about freedom are discussed (and disputed) further in 4.1.

3.4.10 Honderich and life-hopes

Ted Honderich has argued that compatibilists and incompatibilists are concerned with different life-hopes, with compatibilists being satisfied with hopes that are possible to realise under determinism. Meanwhile, incompatibilists have hopes that require indeterminism. The second main claim of the randomness argument is that because indeterminism is what could be characterised as randomness, there is no independent reason to wish for indeterminism as part of freedom aside from desires that amount to just desiring indeterminism for its own sake. Thus, a claim such as Honderich's that there are life-hopes that require indeterminism is a potential place to look for a way to prove the second main claim wrong.

Honderich's two kinds of life-hopes could be summed up as *origination* and *voluntariness*. Voluntariness is just wanting to do things for your own, embraced[232] reasons (2.4.1), and is compatible with determinism. Origination is really a mixture of different incompatibilist requirements discussed elsewhere in this work: ultimate origination, feeling that thinking of the deterministic history of actions destroys the sense of responsibility, and the desire for the future to be open, not determined in advance.[233]

As stated, I discuss these aspects elsewhere. Ultimate origination is discussed below in 3.6, responsibility below in chapters 6, 7, 8 and 9, and the notion of an open future in 2.6.2 and 3.4.11.

Aside from all this, we are left with an argument that people do have these kinds of life hopes, and that matters. Though this does not answer the argument that

[232]On this, see also 8.3.1 in the present work.

[233]Honderich 2002, chapter 8.

concrete control is lost, it could show that there is something valuable enough about indeterminism that it is worth the trade-off of some concrete control. Life-hopes sound like something that is worthwhile. Since we are looking for loopholes in the seemingly solid randomness argument, this could be just such a thing.

This argument does not really work either, however. If we *have* some hopes or other, and our actions are deterministically caused by our mental states such as desires and hopes, then obviously we may be able to fulfil these hopes. Only if these hopes are specifically about indeterminism itself can we say that they cannot possibly be fulfilled under determinism.[234] Honderich's life-hopes of origination may not sound like they are only about wanting indeterminism, but, besides of the specific arguments about ultimate origination, responsibility and an open future I discuss elsewhere, the current line of thought inevitably leads to the conclusion that they are about the denial of determinism. After all, what I said about how, if we have some desires, we could fulfil them under determinism because we have them, is definitely a voluntarist conception in Honderich's scheme, and thus not an incompatibilist one. Thus, if we have hopes or desires that *actually* contradict determinism, then based on this line of thought, they have to involve a desire for indeterminism itself. If we desire something that requires indeterminism, but we do not simply require indeterminism, then we need to come back to asking whether the consequences of indeterminism are desirable – but then the argument about life-hopes can only say that (allegedly) we *do* desire indeterminism (or its consequences), and it will no longer give any reason to say we *should* desire them. If the life-hopes that require indeterminism were to be valuable in such a way that they give a reason to desire indeterminism, they would need to amount to more than a desire for indeterminism. The second main claim that there is no reason to desire indeterminism beyond a desire for indeterminism itself remains unchallenged – and it is interesting to see how many differently expressed incompatibilist desires

[234]I take Honderich to be making a weaker version of this point himself in 2002, p. 130.

continue to canalise into the same place.

We are left with nothing but the idea that indeterminism may be valuable to freedom in the sense that we seem to want it. This foreshadows the discussion in the next chapter (more specifically 4.2, and 4.3) about how incompatibilism and libertarianism are left with little else than some intuitive requirements to recommend them, and how it is almost a contradiction to say one wants something that is enabled only by indeterminism. It also foreshadows the upcoming discussion about how libertarian ultimate origination ends up just meaning indeterminism – section 3.6 below.

3.4.11 Only one alternative

One of Honderich's points in the topic of section 3.4.10 is that if determinism is true, only one alternative in the future is possible for us[235]. This could be seen as threatening to freedom and as an independent reason to desire indeterminism instead. However, on closer examination, it can be seen that it gives no such independent reason.[236]

Suppose the situation under determinism. We focus on a particular event *E* that is racing towards us in the future and that can only happen one particular way. *E* could be a particular choice someone makes, or something else for that matter. Before *E* has happened, it may not be clear to us which event is going to take place at that time. It may be predictable in principle (though see 5.2.2 for why this is less universally true under determinism than one might think), but usually, it is not. What we do know, though, is that only one option is possible – whichever it will be. Furthermore, the conditions now are already such that that one option is the only one that will be realised. However, whatever causes *E* has not really finished causing it

[235]Honderich 2002, p. 94.

[236]This discussion is related to the discussion about determinism and inevitability in section 2.6.2 in the present work, though the point being made is not the same one.

until the moment that *E* happens.

Now, suppose the situation under indeterminism. Some event *E* is coming up and is going to happen, but it is not yet determined which event will be *E*. However, once *E* has happened, then no event contradicting of *E* (e.g. Alice turning left if Alice in fact turned right) is possible, because *E* already happened.[237]

What holds true in both cases is that some event is going to happen, it is only caused to happen once it does happen, we probably do not know what it is with absolute certainty until we get to the time when it happens, and once we know, we also know no other version of events is possible any more. If this is a problem for us in the deterministic scenario, why is it not equally much so in the indeterministic

[237]Cf. Denyer 1981, pp. 42–46, where an argument is given that aims to prove that what is true of the present is true necessarily.

scenario?[238, 239]

Concretely, *E* could be good or bad, desirable or not, under both determinism and indeterminism. It might be in accordance with our wills in either case. Of course, it could only be completely reliably good or in accordance with our desires in the right kind of deterministic scenario.

As with life-hopes above, whatever specific thing you might want *E* to be, that would always be possible under some form of determinism if it were (logically) possible at all. Further, under both determinism and indeterminism, only one of

[238]This point resembles *logical fatalism* but is different from it. Logical fatalism, shortly put, says that only particular future events are possible now in any case because it is now true that these events will happen at the future time when they happen. I think this is meaningless use of language: that *E* happens is true when it happens, and that *E* happens at time *T* is just true (timelessly; and these mean the same thing), but *"It is true at time T_1 that *E* will happen at T_2" has no meaningful truth conditions; how is it true *now* that something is true in the *future*? (Cf. Kokko 2024a.) What my present argument says is not that whatever happens is the only possible thing beforehand, but that, determinism or not, it is the only possible thing afterwards (and beforehand, there is not necessarily much of a difference).

Logical fatalism could be formulated so that it is not tied to time, though. To demonstrate this, consider an interview with Alicia Finch (Finch 2020). Finch explains the idea of logical fatalism, and ties it to time the way I explained above. (In fact, her explanation sounds rather like the consequence argument (see 2.4.1 and 4.1), only without even needing determinism.) However, she also mentions logical fatalism following from the principle of bivalence. The following quote from her might as well be describing logical fatalism as following from the idea that something is true because it is happening right now, and if it is true, it cannot logically be otherwise: "If every proposition is either true or false, and not both, then it seems like no-one's ever able to do anything other than what they actually do, which means that no-one ever acts freely." To be clear, the quote is taken out of context to present a meaning Finch does not intend – she goes on to say that the argument is about what what will be true in the future being true now. However, given the meaninglessness of the idea of something about the future being true now, this is the only sense in which logical fatalism is true: determinism or indeterminism, once you make a choice, you cannot make other contrary choices as well. This presumably does not have the sting of the alleged problem that something is determined in *advance*.

[239]For a similar point, see Dennett 2002, pp. 91–92. See also Dennett 2015, pp. 131–132, where it is tied to the difference (or absence thereof) between randomness and pseudorandomness, and pp. 136–137, 151-152. Also cf. Almeida & Bernstein 2011, p. 485.

mutually exclusive possibilities could be realised at the same time. Only if what we want is literally and exactly for future possibilities to be open – for multiple future scenarios to be (nomologically) possible – before the moment at which future becomes the present can we say that we are losing something desired in determinism being true. However, to say that we want multiple futures to be possible is just to say that we want indeterminism. Thus, the idea of wanting multiple futures to be open does not give any independent reason to want indeterminism aside from wanting indeterminism itself.

It may in fact be a good thing that only one option is realised (if that is the case). Imagine if they all were realised, as in a quantum-mechanical many-worlds scenario where the universe splits into several universes at each seemingly indeterministic event, and all the possible outcomes happen. Then, you would know that whatever you choose, the other options are going to happen elsewhere regardless. In some sense, your choice would then be meaningless, and you would be unable to prevent the options you do not want to happen from happening in some futures.[240]

3.4.12 Appeals to the value of responsibility (why they are not discussed yet)

It needs to be mentioned at this point that it could be argued that libertarian free will gives us something valuable in terms of making moral responsibility at all possible. Since I am postponing discussion of responsibility until chapter 6, I will also put off answering any such claims until then. (See especially 6.3) I will, however, later show that the kind of concept of free will I end up defending for other reasons is also a good match for the kind of responsibility that is worth wanting – see chapters 8 and 9.

[240]Slattery 2014, pp. 133–134.

After all of this has been said and all these arguments refuted, is there no loophole in the randomness argument? There is a partial one, which we turn to next.

3.5 Uses of randomness

> Normally noise is the enemy of information, but it can be the friend of freedom and creativity.
>
> –Robert O. Doyle[241]

The randomness argument states that indeterminism always leads to a decrease of concrete control, in principle and insofar as it makes a difference. I spent the previous section showing the different ways in which randomness cannot be avoided, and that it does lead to a decrease of concrete control. There is, however, one thing close to a loophole in this argument, and it stands apart from the other suggestions enough to deserve its own section. In this section, I go outside the box and ask whether randomness might not increase concrete control in some circumstances. This leads to rather different perspectives than those presented above.

3.5.1 Randomly avoiding worse options

First, a simple idea. If I were to ask the question, "In what situation would it be desirable for a choice to be undetermined rather than determined for practical rather than *a priori* reasons?," the most obvious answer is "When the choice made is determined to be undesirable." Undesirability might mean, say, that the agent will regret that choice later. All that we need to postulate is that it would be less desirable than one of its alternatives. In such a case, we could say that the chance of randomly choosing the more desirable alternative would be better than being determined to choose the less desirable alternative.

[241]Doyle 2013, p. 246.

Refuting this with the idea behind the randomness argument is not very complicated. It might be more desirable in the individual case to be undetermined between making the good choice or the bad choice than to be determined to make the bad choice, but by the same valuation of the alternatives, it would be more desirable than either to be determined to make the good choice, with no chance of making the bad. Applying the same idea further, if you wanted indeterminism to apply only in the cases where you are about to make the bad choice, you would need to have a reliable and thus more or less deterministic mechanism for choosing those cases when you are about to make the bad choice – and, again, it would be better to go on from there to deterministically avoid the bad choices. Finally, if you wanted the indeterminism to apply at all times in order to avoid bad choices, that would only work to produce desirable choices more often in case you were so bad at making the desirable choices that you would do worse than chance. In that case, it would again help even more if you got better at making choices such that you got the desirable ones more often than chance.

In sum, having a chance of randomly making a different choice would be an improvement in desirability in an individual case compared to being determined to make the less desirable choice (by whatever standards of desirability[242]), but being determined to make the more desirable choice would always be even better. Hence, indeterminism cannot offer anything in this sense that is not worse than what *some* form of determinism can offer.[243]

3.5.2 The finiteness argument

A more interesting idea relates to the two-stage models (3.4.6). Might it be useful to randomly generate a list of options to make a choice from? What if we need to get a

[242]The question of what kind of choices and mechanisms for making them should actually be regarded as desirable will be addressed below in 8.3.

[243]Chapter 5 discusses how not just any form of determinism will not do for freedom.

randomly selected list of alternatives to choose from because it would take too much effort to work out the best of all alternatives that might possibly occur to us? Would that not mean that randomness would be doing some useful work, and thus that indeterminism might help, not hurt in this part of the process? Thus, there could be a motive for adding indeterminism.

Robert O. Doyle suggests something similar in his treatment of two-stage models[244], complete with a loose account in scientific terms of how such a thing would happen in practice. His focus is slightly different; he points out that randomly generating different new alternatives for consideration is a useful form of creativity. I agree that this could be concretely useful. At the same time, though, Doyle is quite explicit that some suitable kind of indeterminism is a primitive requirement for freedom. Putting that last point aside, he still gives an independent motivation for indeterminism in two-stage models. We could put this by saying that we get a useful, creative selection out of the vast number of possibilities that we could think of.

Thinking narrowly, the randomness argument counters this quickly. As above, *some* form of determinism still gives more concrete control. There is some potential selection of alternatives from among which the random selection of a small range of alternatives for the agent's more considered selection happens. If the agent *were* able to choose between all those initial options intelligently, they would have more concrete control than if the choice was random. Hence, the right kind of determinism is still better for concrete control in principle.

However, suppose the situation in the actual world is this: We are too finite and limited to make a determined choice between all possible alternatives we could potentially think of.[245] This is plausible since there seem to be potentially infinite options in any given case. In such a case, we cannot possibly reach the ideal of

[244]Doyle 2013.

[245]This is something of a flipside to the argument in 2.6.3 that a vast universe contains enough material to effectively support novelty and creativity even under determinism.

deterministically choosing between all possible alternatives and having maximal concrete control. We will need some reasonable compromise, and a random selection of options to take into closer consideration can be a helpful one. The creativity aspect that Doyle speaks of could also be helpful.

This *is* a loophole in the randomness argument, as I predicted in the beginning of the previous section (3.4). Just as I said there, it may be that, considering enough concrete details, it may turn out that the logically inescapable argument does not cover everything. A certain kind of determinism would still be more ideal in terms of concrete control in principle, but what use is that if such a thing could never be reached?

This line of thought, let us call it the ***"finiteness needs randomness" argument*** or ***finiteness argument*** for short, establishes that randomness can indeed be useful or even necessary for concrete control in the right circumstances. Nevertheless, it does not establish that indeterminism is necessary, nor does it serve to validate the other original motives of those thinking so.

The first point against using the finiteness argument to defend incompatibilism is simple and decisive. While establishing a use for *randomness*, it does nothing to establish a need for *indeterminism on the ultimate level*. There is no reason why higher-level pseudorandomness (see 3.2.2) could not do the work of non-reasoned selection of the pool of options. The finiteness argument does not disprove compatibilism.

Secondly, though relatedly, the finiteness argument seems to have very little to do with most of the intuitions motivating the incompatibilists and libertarians in the first place. They sometimes speak of related points as discussed especially under 2.6.3, but arguments related to origination (2.4.1), alternative possibilities (2.4.2) and antireductionism (see 2.1.6) are strictly separate from this, even if some connections may be found.

Does anyone's intuition that we need to be able to do otherwise to be free and responsible in any way suggest that this need is because we can sense that otherwise, we would be overwhelmed by having to actually decide between too many options?

This is certainly not a motivation most incompatibilists mention. Indeed, the intuitively plausible-seeming theory of choosing indeterministically between alternatives that are present in the mind (3.4.5) is the complete opposite of this picture. The requirement of ultimate origination also does not seem to be motivated by any considerations of needing to choose a pool of options randomly. Hence, though we arrived at the finiteness argument via the two-stage models that are motivated by some of the common intuitions behind incompatibilism, it seems we are now in completely different territory, and not about to validate these intuitions. (This is unsurprising remembering that the theme of this section is to find different kinds of motivations for seeing randomness as useful for freedom.)

Thirdly, there is the point about how the randomness argument could technically answer this point by pointing out that being determined to pick the best option would still be better. Even though I argued above that it is not enough of an answer to say this, it is still relevant in a roundabout way. Incompatibilism, unless we adopt a weaker form of it than has been discussed previously, is about how determinism is in conflict with determinism *in principle*. If determinism would still be better in an ideal situation, even if that ideal is unreachable, that speaks against the idea that indeterminism is an essential part of freedom in principle. Since the finiteness argument only says that randomness can be useful or necessary because we are finite beings, even though a being capable of perfect free decision-making[246] would be just fine being determined the right way, it does not prove that indeterminism is incompatible with free will in principle by the nature of either concept. Thus, even if pseudorandomness was not enough, the finiteness argument would not prove that it is in the nature of free will to be incompatible with determinism.

In sum, the "finiteness needs randomness" argument does show that randomness can be useful for concrete control. However, it seems to have very little

[246]Perhaps Reid's God; see 11.9.

to do with most original incompatibilist motivations, and it does not serve to validate actual incompatibilism because pseudorandomness works just as well as indeterminism on the ultimate level. Either way, we may need randomness to generate options to choose from, but if we did not need it for this reason, there is still no reason to think we need indeterminism for freedom other than if we take it as a premise that we do. Thus, for example, I have no objection to Doyle's two-stage model as a possible model for freedom, but that does not mean I have to endorse the motivation of it that (true) indeterminism is required. Indeed, in light of this model, it would be rather peculiar if it turned out that the model is otherwise true except that the randomness involved in generating options is chaotic pseudorandomness – and we would have to conclude that, in spite of that model working exactly as it had before otherwise, we are suddenly not free (cf. 2.5).

3.5.3 Choosing without sufficient reason

Putting the finiteness argument aside, another similar use for randomness is found in "Buridan's Ass" scenarios. Buridan's ass is a hypothetical hungry donkey that is (in one version) situated exactly halfway between two piles of hay. Both are equally desirable and equally easy to get to, meaning that the donkey has no reason to choose one over the other. If it could only act on a sufficient reason to choose one way or the other, it would starve due to not being able to decide.

Clearly for the donkey to or anyone in the same sort of situation would need to make a choice that would be random on that level of description on which it is true that there is no reason to choose one over the other, or else make no choice.

However, there is still no reason why the choice would have to be indeterministic on the ultimate level. There might be a deterministic mechanism – even deterministic on the psychological level – that would lead to choosing one option even if there was no reason to choose that option rather than the other. In a simplified example, the donkey might have no reason whatsoever to choose the right pile instead of the left, yet it might be neurologically programmed to always choose

the one on the right side in such cases. The piles of hay would not be identical with respect to which one is on the right and which one is on the left, and while this difference might give no reason to choose one or the other, it could make a difference for the mechanism actually operating within the agent's choice. If there was no difference whatsoever between the two choices, even on a physical level, so that for example the piles of hay would also have an identical location, then there would only be one choice (pile), not two.

Thus, in "Buridan's Ass" scenarios, indeterminism of a certain sort is useful or even necessary for reasons-responsiveness because without it, neither choice can be made. However, this is a very limited sort of indeterminism that only needs to apply on a level describing the agent's reasons and not on any other level. In fact, it is not *possible* to formulate these scenarios so that they would require indeterminism on the ultimate level, because if there is no difference between the options on the ultimate level, there is no difference that would make them two distinct options – in fact, no possible difference at all, since the ultimate level is postulated to be one on which everything else is based. Similarly to the finiteness argument, the "Buridan's Ass" scenarios neither require indeterminism on the ultimate level, nor do they have anything to do with the usual original motivations of incompatibilism.[247]

As an overall conclusion to this section, there are a few uses for randomness even from the point of view of concrete control, or at least from similarly practical considerations, but these go in a different direction from the demands of most incompatibilists, do not require indeterminism on the ultimate level, and do not prove that any kind of randomness is needed in principle in the ideal case.[248]

[247] Alan Donagan argues that flipping a coin to decide in such cases is taking up an extrinsic cause, which can be rational (Donagan 1987, pp. 152–153). Cf. Dennett, 2015, p. 76.

[248] An additional advantage of randomness might be that one's actions not be too predictable to enemies and rivals, which of course can be handled with pseudorandomness as well (Dennett 2015, pp. 72–74), unless one manages to pick a fight with Laplace's Demon.

3.6 Ultimate origination and indeterminism

As has been pointed out before (2.4.1), libertarians are likely to want the agent to be the ultimate origin of their own actions, insofar as they do not want something else to be a more ultimate cause.[249] The usual problems about the compatibility problem, the intelligibility problem, and the randomness argument affect the very notion of ultimate origination heavily. It seems as though ultimate origination requires both concrete control and indeterminism at the same time.[250]

The problem is this: It seems that if you are the origin of your choice, then you must make that choice for your own reasons. However, for you to be the *ultimate* origin of your choice, it seems that the choice must not be predetermined by anything before or outside you. How can anything be both? To make a choice that is really determined by you, it seems, you must have concrete control, and you must make the choice for reasons. Thus, before you can make any choice, you must already have some kind of properties. But then, you cannot have chosen those properties yourself, because that kind of choice would already have required you to have some properties.

So, if indeterminism contradicts concrete control, and determinism contradicts "ultimate origination", then there is no such thing as "ultimate origination". However, we can be generous (and I can spare myself having to use quotation marks

[249]Technically speaking, the definition of *libertarianism* I have introduced for use in this work (1.3) means that libertarianism is about indeterminism and not about ultimate origination, so that anyone calling themselves a "libertarian" but demanding ultimate origination and not indeterminism would not count as a libertarian here. This is one reason I do not follow Saul Smilansky in speaking of "libertarian free will" below (6.2) but instead say "ultimate origination". (Other reasons include consistency in my own terminology and not wanting to, as it were, grant the concept as being usable the way Smilansky uses it.) However, since the argument in the present section is that ultimate origination as usually understood by views typically labelled as "libertarian" implies indeterminism in any case, there should henceforth be no problem in grouping ultimate origination views as libertarian.

[250]For sources discussing this or related problems, see e.g. O'Connor 2011, pp. 320–321; Wiggins 1973, p. 47; Slattery 2014, chapters 23–25.

all the time) by conceding to speak of "ultimate origination" without requiring concrete control. Thus, the expression is used in this dissertation to mean *being the ultimate origin of one's choices in a libertarian sense*.[251] This means that it has two requirements for an agent to be the ultimate origin of their choice: one, the choice is the agent's own in some sense that does not require concrete control, and two, the choice, or something suitably leading up to it, is or was indeterministic. Meanwhile, I will reserve the term *full ultimate origination* for the (impossible) version that requires control in the uncompromised sense.[252]

What does this leave us with? Ultimate origination implies indeterminism. It also implies some form of authorship or ownership of one's choices or actions, but as we have seen above (3.4.1 and 3.4.2), that makes no difference for the randomness argument. From the perspective of main claim 1 ("indeterminism contradicts concrete control"), the randomness argument applies to ultimate origination in exactly the same way as it does to the demand for indeterminism. In fact, the demand for ultimate origination pretty much *is* the demand for indeterminism – coupled with the demand for some kind of authorship that the libertarian demand is usually paired with anyway.

What about main claim 2, "There is nothing independently desirable about indeterminism in choices"? The way ultimate origination is seen, it looks as though it is an independently desirable thing to be gained from indeterminism, from cutting the chain of determination. However, there is no unquestionable form of ultimate origination fulfilling both the requirements of control and of there being no more ultimate origin of your choices. The form of ultimate origination that we ended up talking about is a libertarian compromise that accepts compromising concrete control

[251]Of course, hard incompatibilists would probably say being the ultimate origin in a *real* "libertarian" sense is impossible, for the same reasons as I say in this section that full ultimate origination is impossible (see e.g. Smilansky (e.g. 2000), who is discussed in 6.2).

[252]This will be of use especially when discussing Saul Smilansky's view, mainly in chapter 6.

but does not accept compromises about avoiding predetermination. Thus, the second main claim also applies to ultimate origination because the concrete consequences of ultimate origination are the same as the consequences of the form of indeterminism accompanying it. There is nothing more to be concretely gained from it any more than there is from any of the specific libertarian theories where authorship is postulated without concrete control. A person might have desires aimed at the notion of ultimate origination rather than that of indeterminism, but if indeterminism is required for ultimate origination, the demands amount to the same thing, with only different mental labelling.

In summary, the demand for ultimate origination as a condition of freedom is, for current purposes, identical with the demand for (some form of) indeterminism. All you get from it is the cutting of the deterministic chain of events that would determine your choice if it existed. A desire for it amounts to a desire for indeterminism, as there is nothing else to be gained from it.

The desire or demand for ultimate origination can also be seen as part of a strange, hyper-individualistic way thinking, as opposed to a view that better sees and accepts connections between the agent and the rest of the universe. This theme will be discussed later in section 9.1.2.

3.7 Conclusion to chapter 3

This chapter has dealt with the randomness argument, which can be summed up, among other ways, as follows: Indeterminism as applied to choices only means useless randomness and loss of control, and for that reason, it can also offer nothing independently desirable. Through a thorough discussion of different aspects of the argument, and then of many possible counterarguments, I have shown that this applies in principle no matter how the indeterminism is treated, where it is placed in the process of choice, and how small its role is made. If indeterminism makes any difference at all, it only grants a loss of concrete control. The sole exceptions were cases where it may well be useful to be random, and even in those cases,

pseudorandomness would work just as well. The demand for ultimate origination also amounts to nothing more than a demand for indeterminism.

It seems that common intuitions want both indeterminism and concrete control, but these demands are at odds. In the next chapter, I will look further into this and continue the argument of this chapter to make a conclusion about what we should think about the relationship between freedom, determinism and indeterminism.

4 From the Randomness Argument to Compatibilism

In this chapter, I continue the argument from the previous chapter to one of its logical corollaries: if indeterminism always compromises an important aspect of freedom (main claim 1 of the randomness argument), and indeterminism is not independently desirable for free will (main claim 2), it makes sense to simply discard the requirement of indeterminism and thereby embrace compatibilism. Even if this is not in accordance with all our intuitions, it may be the best we can get, and the price to pay seems to have been revealed as illusory.

I start by taking a closer look at one perspective that opposes the one I am going to propose: Peter van Inwagen's argument for incompatibilist free will in his *An Essay on Free Will*.[253] This argument is based largely on the consequence argument and related intuitions and trains of thought. Van Inwagen admits that his argument ultimately hinges on a certain key assumption; I will show that while this assumption is intuitively strong out of context, considering all the factors of the situation leads to our needing to make a choice, one going beyond puzzling over just how we feel about it.

After this, I launch into my own positive argument. I start by showing that the requirement for indeterminism and the requirement for concrete control bifurcate the logically possible definitions of free will into two camps, one based on indeterminism and one on determinism. After we see what these options are, without any conflating and in truth incoherent "third options", we can compare what they

[253]Van Inwagen 1983.

have to offer and make a reasoned choice between them. I also discuss the possibility and possible usefulness of choosing to say that since we cannot combine all common requirements for free will in one concept, we should embrace hard incompatibilism instead.

4.1 The randomness argument versus van Inwagen's consequence argument

Different sides of Peter van Inwagen's views have been explored briefly before.[254] In this section, we will be concerned with his formulation of the consequence argument. As discussed in 2.4.1, the consequence argument states that since under determinism, our choices are determined by states of affairs before we were born, and those states of affairs are not up to us, or we have no choice about them, then under determinism our choices are not up to us (or we have no choice) either.

If you dig a little deeper into this argument, you will find the incompatibilism built in into one of its premises – albeit one that seems intuitively correct enough. As van Inwagen admits[255], his consequence argument depends on the rule of inference *"β"* that has the following content:

> If it follows from *X* that *Y* will happen, and no-one has or ever had a choice about whether *Y* will follow from *X*,
> And *X* is true and nobody has a choice about whether *X* is true,
> then nobody has a choice about whether *Y*.

β is also called the *transfer principle*. We could concede that if nobody has a choice about *X*, and *X* applies to everything as per universal determinism, then free

[254]Besides of what is referenced next, see 3.3.1, 3.4.4 and 3.4.9.

[255]Van Inwagen 1983, pp. 94–97.

will is impossible.

Should we accept β? It is begging the question against compatibilism, since to say compatibilism is true is equivalent to saying β does not apply in the case of universal determinism. However, the argument is that β can be established independently.

A part of it is that β is supposed to be independently intuitive. Of course, it is to some extent. Van Inwagen tries to argue that it is that by giving examples where everyone can agree that β must be right. (While there is seemingly nothing literally everyone could agree about, these examples do push the believability of disagreeing with them to an extreme in my judgement.)

> Alice has asthma and no one has, or ever had, any choice about whether she has asthma;
>
> If Alice has asthma, she sometimes has difficulty breathing, and no one has, or ever had, any choice about whether, if she has asthma, she sometimes has difficulty breathing;
> *hence,*
> Alice sometimes has difficulty breathing, and no one has, or ever had, any choice about whether Alice sometimes has difficulty breathing.
>
> The sun will explode in 2000 AD, and no one has, or ever had, any choice about whether the sun will explode in 2000 AD.
>
> If the sun explodes in 2000 AD, all life on Earth will end in 2000 AD, and no one has, or ever had, any choice about whether, if the sun explodes in 2000 AD, all life on Earth will end in 2000 AD;
> *hence,*
> All life on Earth will end in 2000 AD, and no one has, or ever had, any choice about

whether all life on Earth will end in 2000 AD.[256]

Van Inwagen also effectively challenges compatibilists to provide examples where β seems wrong and compatibilism is not assumed, since his examples do not assume incompatibilism.[257] I freely confess that it seems to me that there are no such examples to be had.

The flaw in this is that van Inwagen's examples are all ones of events where agency or choice is impossible quite regardless of questions of determinism and free will. There is no agent making a putative choice. Van Inwagen's position is just as constitutionally unable to provide examples of β that assume neither compatibilism or incompatibilism *if it would actually come up*. His using these examples in some sense does assume incompatibilism, because he presents them as analogous to the case where there *is* an agent making a putative choice in a deterministic world. That is, by asserting the validity of the analogy, he assumes that what holds for cases where there is no agent and thus tautologically no choice also holds for cases where there is a putative choice under indeterminism.

This means that the real question is whether β holds in cases where there is an agent making a putative choice – whether these are analogous to the situations with no agent or not.

As a note connecting to a future topic, the lesson of this might instead of anything else that has been said be that we have different ways of looking at things under natural, mechanistic events and under agency. This idea will be extensively explored in 5.3. Insofar as the puzzle about the consequence argument is that it seems wrong but β also seems right, that discussion might be the best answer to it.

But back to evaluating β and the consequence argument. We are left with making a choice about whether to accept β. We should not forget what happens if we

[256]Van Inwagen 1983, p. 98; formatting altered slightly, italics in original.

[257]Van Inwagen 1983, pp. 101–102.

do accept it. Given the randomness argument's main claim 1, which van Inwagen has not countered (see 3.3.1), indeterminism will cost us a vital part of freedom – at least insofar as it has an effect. Recall that van Inwagen requires that all possibilities that a choice is being made between must be possible (see 3.4.9). Thus, we are stuck with at least all possibilities we have thought of being possible such that we might randomly choose any of them. Given this, the only way we can mitigate the randomness is if we say that options we realise are bad are highly unlikely to be chosen – but if van Inwagen thinks the possibilities really need to be open, it makes little sense to interpret his theory so that some of them are vanishingly unlikely. Thus, if we do accept β, we are left with a necessary requirement for freedom that itself makes freedom pretty impossible. This leaves us with either hard incompatibilism, accepting β and biting the bullet on the loss of concrete control in favour of randomness, or biting the bullet on rejecting β for agential action.

In the end, the consequence argument has not proven anything more than we already granted: there are some intuitions that make it seem as though free will contradicts determinism, taking various forms but all suggesting the same proposition. At the same time, indeterminism is doing no better, not even in terms of intuitiveness. At this point, there is little we can do other than to accept hard incompatibilism – or take one or the other horn of the dilemma.

So, if we formulate two different conceptions of free will, each based on one side of the contradictory requirements, what do we get?

4.2 Deterministic and indeterministic models of free will

In this section, I will argue that our various intuitions point not to one definition of free will but two irreconcilable ones. I will sketch these definitions out in forms that finally avoid any internal contradictions.

4.2.1 The intuitive contradiction

We have seen that common intuitions can be found supporting several different properties as being necessary for freedom of the will. Alternative possibilities and ultimate origination are apparently needed on one hand, but on the other, predictability and the agent's control also seem essential. The randomness argument shows absolute alternatives and practical control cannot be had at the same time, whereas (full) ultimate origination seems to embody both in a way that leads to a contradiction within the concept itself.

At this point, we need to admit that these intuitions that we have been following are contradictory. It is not just a matter of contradicting intuitions between incompatibilists and compatibilists. As described in 2.2.1, studies give contradictory intuitions held by non-philosophical individuals surveyed, and individual incompatibilists themselves seem to show signs of something similar (see 2.2.2 and 4.3.1). As a compatibilist, I do not think my positions stems from having very different root intuitions from these incompatibilists.[258] We come back to generalising about intuitions that people in general often seem to have, and now, I generalise as follows: common intuitions about what free will implies and requires seem to be self-contradictory, in the sense that different intuitions held by the same people tend to contradict each other.

This means that we cannot have a model of free will that fulfils all the common intuitive requirements. Such a model would be self-contradictory. What we can try to do is construct models each bringing together properties of freedom that can be combined. As it turns out, less than surprisingly, we can have one such compromise model for compatibilism (or soft determinism) and one for libertarianism.[259]

[258] I would certainly not spend all this effort to defend one intuition over another.

[259] 'Trick Slattery (2004, e.g. chapter 3) makes much the same case, though he aims to affirm hard incompatibilism.

4.2.2 The compatibilist compromise: Non-ultimate control

The first of the two possible compromises involves dropping the requirement of the possibility of doing otherwise in the strictest sense as well as that of ultimate origination. What we are left with is that freedom involves being the proximate origin[260] of your own choices, and probably being able to do otherwise in some less extreme sense, on a higher level of description. It also, importantly, involves concrete control. Thus, this conception of free will involves your choices being determined by your own reasons.

Now, perhaps not just any reasons will do. Maybe some of the reasons or motives affecting an agent are less free than others, perhaps somehow less genuine. This was already discussed in section 2.6.1, and it will be discussed further in sections 5.4 and 8.3 (and see also 9.6.1); the questions will be bypassed here with this mention that it is not necessarily the case that all reasons or especially motives are "free ones".

That aside, what we are left with is the picture that our free choices are ones where we have concrete control over them and where we are their proximate origins. Because they are this, they cannot be indeterministic at the same time, and in this conception, that would only be a hindrance anyway, as it would compromise concrete control. Our choices need not be determined by anything outside of us except in the sense that events originally outside of us created us. They are also not caused compellingly by any single factor that overrides the rest of our motives. We can trust that our choices are good ones for ourselves if they are free in this sense, and we do not need to fear what they will be like. The price we pay for this is not being able to include indeterminism from our shopping list of intuitive requirements.

[260]On being a proximate origin cf. e.g. Vargas 2013, pp. 7–8, Balaguer 2014, p. 47 (though that last is part of an argument about voluntarism, which, though I do not really discuss it here, I see as insufficient as an explanation of free will; see 8.3.1 for a related point).

4.2.3 The libertarian compromise: Agent-causality or postulated ownership

If we take the other horn of the dilemma, we are left with the option having indeterminism as part of our choices, at least some of them, at the expense of concrete control. The indeterminism is seen as a necessary condition for the choice being free, and as compatible with authorship.

We can immediately say that this conception of freedom achieves some of its negative goals, avoiding certain things: every choice is not determined in advance, and events in the distant or near past do not lead to or determine the choice. It is, definitionally, possible that the chooser choose more than one thing, until after the choice is actually made.

On the problematic side, free choice in this model (at least on the times it is indeterministic, which I will not repeat again in this section) lacks concrete control and contains an element of randomness that acts as a discontinuity between the chooser's reasons and motives and the final choice. To the extent that the choice is indeterministic, it does not stem from anything concrete in the agent. This also leads to it lacking concrete control, and the agent may have to worry about what decisions they will make in the future, not because those decisions might not be free, but precisely because they might *be* free in the sense talked about in this conception.

As we have seen (3.4.1 and 3.4.2), libertarians can insist that the agent can still be in control of their own choices – not in the sense of concrete control, but some other sense, one in which they *just are* causally responsible, even though the observable features of the situation, detached as they are from the agent's motives, are just the same as if the choice was made by an outside source of randomness.[261]

[261]It could be postulated that the observable features of the situation would be different from outside influence in that the indeterministic free choice would *feel* like a free choice, and the outside influence would not. However, the features of the situation would still be the same as in a postulated situation in which an outside random influence determines your choice but you feel as if it is your own choice that you are making. The question here is the outcome of your choice and its connection to the agent's own motives, not what things feel like. Such a feeling could be an illusion, and in this postulated case, it would be that at least in a significant sense.

Further, libertarians may be willing to accept that this kind of choice also allows one to be the ultimate origin of one's own choices. Again, the negative side of this claim may be true: there is no more ultimate source than the agent. Meanwhile, the positive side is more questionable: the person just is the origin of the choice, even though nothing in the situation makes it so.

What we end up with is what I call *postulated ownership*: the agent is said to be the origin of their choices, even though, insofar as the choices are indeterministic, an analysis of the situation attending to the relevant factors – the connection between motives and other features of the agent – seems to show otherwise. We could also broadly call this agent-causality, since agent-causality typically ends up saying the agent *just did cause* their choice, and it is a more familiar term. If the determinist-compatibilist conception described above had to compromise by giving up indeterminism, the indeterministic-libertarian one has to compromise by settling for postulated ownership without (and instead of) concrete control.[262]

A point to this effect was also made by Chisholm when introducing agent-causality:

> The only answer, I think, can be this: that the difference between the man's causing *A*, on the one hand, and the event *A* just happening, on the other, lies in the fact that, in the first case but not the second, the Event *A* was caused and was caused by the man. There was a brain event *A*; the agent did, in fact, cause the brain event; but there was nothing that he did to cause it. This answer may not entirely satisfy and it will be likely to provoke the following question: "But what are you really adding to the assertion that *A* happened when you utter the words 'The agent caused A to happen'?"[263]

---

[262]For a definition of free will in these lines, see also Balaguer 2014, p. 129.

[263]Chisholm 2002, pp. 54–55. Chisholm goes on to say that this is a problem faced in explaining causation generally, not only agent causation. I am not trying to answer questions of causation here, though I will say it will probably be easier to do so in a

It should be noted that postulated ownership, or indeed agent causality in a broad sense[264], does not in itself imply indeterminism. They could just as well coexist with determinism – and with concrete control, though that would be somewhat redundant, since concrete control already gives ownership without postulation. As for postulated ownership, if we can postulate that a choice just does belong to an agent when there is nothing concrete making a difference between it being their free act and it being an uncontrolled event emerging from their body or brain, then why could we not just as well postulate that it is theirs when they have concrete control over it? As for agent causality, if some acts just are caused by the agent as a "substance", that does not in itself mean that the agent is not acting by deterministic rules that determine which acts it causes based on the external situation.

What determinism contradicts is, again, ultimate origination.[265] Postulated ownership does not require indeterminism (since it does not require anything), but it allows it (if we so choose). The libertarian compromise thus involves a combination of indeterminism and postulated ownership.

I do think it is dubious whether agent-causation in this indeterministic sense is a meaningful concept. If there is no difference between an uncaused random event being part of an agent's deliberative process and a random event caused by the agent being such a part – that is to say, no difference that can be explained through anything

---

case where regularities or some other connection can be observed between the cause and the effect than in an indeterministic case where the result is indistinguishable from a chance event caused by something else or nothing at all. Putting the question of causation aside, the argument I am here putting forward is that it is easier to understand how there can be a connection between the agent and their choice if the connection between the agent's motives and their choice is not purposefully severed by indeterminism. The reasons for this hold even if we have no concept of causation whatsoever.

[264]On this, see also 11.9.

[265]If postulated ownership is enough to make an act yours, we might also have defined "ultimate origination" so that postulated ownership is enough for it, but then we would need another term for what incompatibilists want.

else – then it does not mean anything to say that one rather than the other is true. However, I am willing to grant this point for the sake of the argument. Let us say this is a coherent indeterministic, libertarian view of freedom and agency. If it is one, it also seems to be the only one.[266]

What matters to me and the current argument is that anyone endorsing such a view will now have to bite the bullet on the randomness argument. In this model, we have agent-causal control, we even have ultimate origination (if only in the negative sense we defined), but we do not have the kind of control that avoids randomness. Thus, both of our two models really are compromises in respect to the desiderata introduced so far. To put it simply, you can have concrete control or you can have ultimate origination and strict alternatives, but you cannot have both.

The next step is to figure out what we should think about all this. Do we conclude that hard incompatibilism is correct, or can we somehow reason that one or the other of the two mutually exclusive options should be regarded as the "right" one?

4.3 Making a choice: Avoiding hard incompatibilism

If we accept the idea that free will is what we collectively intuitively think of it as being, we would now have to admit that free will is impossible at least in some sense. If neither determinism nor indeterminism is compatible with freedom (and there is no third option, as there is not), then hard incompatibilism is true. Fortunately, I do not think that way. As I pointed out in 3.1, I am looking for the "real" definition of free will in the sense that I want to talk about free will in the sense that really matters to us as human beings. If I can show that only one of the compromises above compromises anything that is important to us in our lives, as opposed to merely being a compromise in regards to our intuitions, then I will have reasons to support the one

[266]Of course, it allows for further more specific views about when this kind of indeterministic choice happens – see 3.3.6.

position as “right” without it being intuitively just right.

4.3.1 Determinism envy

There seem to be two main strategies that libertarians use against the randomness argument (not counting ignoring it). The first is to conflate the question of concrete control with that of “ownership”, and go for postulated ownership, not countering the randomness argument. The second, sometimes employed at the same time because the issues are linked, is to make the role of indeterminism as small as possible.

Consider Kane’s model of self-forming actions[267] (see section 3.4.7). In this model, indeterminism is vital for freedom, but most choices do not need to be indeterministic. In those that are, the only options that may be chosen are those that you already have a motive for. And this only applies in cases where there is more than one motive that is about equally strong. Further, it is emphasised that it is important that your choices follow from your own motives.

The question is this: Is this not a case for everything being made to go as much as possible as it would under determinism while affirming indeterminism? Further, the reason it fails to counter the randomness argument is still due to the presence of what indeterminism there is.

If people like Kane believe that indeterminism is a vital requirement for free will, why do they try to get rid of it? Of course, the answer with Kane would be that he wants to solve both the compatibility problem and the intelligibility problem (see 1.1). Nevertheless, he seems to be doing this by invoking determinism to affect what happens in the scenario, and by invoking indeterminism to be indeterminism while affecting as little as possible.

If incompatibilists really want indeterminism, why do they want none of its actual consequences?

[267]Kane 2002a, Kane 2011.

4.3.2 Taking stock: What determinism and indeterminism make possible

To make the decision of which internally coherent view of free will we should adopt – the libertarian–agent causal view or the compatibilist–concrete control view – we will next examine the consequences of each view for different things that can be asked of free will. I will refer to these views below, for short, as what is possible under indeterminism and determinism. Each of the statements below will be followed by a reference, in bold, to the section(s) where that question was discussed.

The following two points are important to remember when reading what follows, as many of the statements would be false otherwise – and since these qualifications are not repeated at every point, it can easily sound as though a different interpretation is meant.

- When I speak of what is possible under determinism, I do not mean that determinism automatically makes it possible, just that some form of determinism allows it.[268] This is because incompatibilism claims that *any* kind of determinism contradicts free will, whereas compatibilism can say that (only) the right kind of determinism is compatible with free will.
- In the case of indeterminism, if I say something is *not* possible, I mean that insofar as indeterminism makes a difference, it makes the thing in question impossible – there can be cases of indeterminism where it has so little effect the result is the same as under determinism. So, in effect, indeterminism that is virtually determinism counts as determinism below. This is because, a point I have already made several times, wanting indeterminism that acts just like determinism is no reason to want indeterminism over determinism (see especially 3.3.5 and 4.3.1).

With that said, here is the evaluation of what is possible under determinism

[268]Cf. Vargas 2013, pp. 38–39.

and indeterminism:

Can we be free of special determinisms, such as biological determinism, and irresistible desires, and otherwise resist temptations? Under both determinism and indeterminism, yes. **2.6.1**

Can we be creative, spontaneous, surprising? Under both determinism and indeterminism, yes. **2.6.3, 3.5.2**

Can we act without being paralysed in "Buridan's ass" scenarios? Under both determinism and indeterminism, yes. **3.5.3**

Do we have alternative possibilities? Under indeterminism, yes, in the strictest possible sense. Under determinism, not in that sense, but on a higher level yes, and in such a way that it is useful, yes. **2.4.2 (also 5.4.4 later)**

Can we deliberate, and can we believe that we can deliberate? Under determinism, yes, in the normal sense that we can weigh different options and come to a reasoned conclusion. Under indeterminism, not in the normal sense, but only insofar as what we would want would be to spend some time thinking about the different options and then endorse any one of them over the others for no reason, though caused by you as a substance. So: under determinism, yes, under indeterminism, no. **3.4.9 (also 5.4.4 later)**

Can we trust that we will do what will be the right choice in the circumstances in the future by our best judgement? Under determinism, yes. Under indeterminism, no. **3.3.4**

Can we control ourselves? Under determinism, yes, and in the robust and useful sense of concrete control. Under indeterminism, yes, but only in the dubious and fairly vacuous sense of postulated ownership, which we also could have postulated for determinism if we chose to do so. **3.3.4, 3.4.2**

Given the previous points, can we be rational agents? Under determinism, yes. Under indeterminism, this is compromised.

Can we have "ultimate origination"? Under indeterminism, only if we accept postulated ownership for undetermined actions lacking concrete control (and thus not meaningfully originated by us). Under determinism, we could have it for an

equally large compromise, by calling it ultimate origination when we are the meaningful proximate origins of our actions but not really their ultimate origins. The most objective evaluation would be to say (full) ultimate origination simply cannot be had. **3.6**

Can we be the proximate origins of our own choices – so that in the here and now, as beings within the world who already have motives, we can follow those motives in that world? Under determinism, yes, without qualification. Under indeterminism, only partly.

Can we be morally responsible? This large question will be discussed starting in the next chapter, but the answer I ultimately give will be similar to what I have said at some other points here: We can be morally responsible in a limited but meaningful way under determinism, and under indeterminism, in a limited, intuition-based, not very meaningful or useful way.

Now, there are some questions remaining that cannot be answered in such an analytical way. Can we be virtuous? Can we take credit of our work as its creator? Do we have the kind of authorship of aspects of our life that is enough for us to feel that we have it? These kind of questions are really something that follows from our choice of a concept of freedom, and further, making the choice on such an analytical basis may not be enough to satisfy our need to feel that we have the right kind of authorship. About these questions, I can try to paint a picture instead of analysing them.

Under indeterminism, we can have meaningful authorship in the sense that the rest of the universe does not interfere upon it. We can be separate from other influences, even if that is at the price of being somewhat separate from ourselves too. Our individuality (in the negative sense of separateness) is fundamental over anything else.

Under determinism, we can have meaningful authorship under a conception where we are a part of the universe. We are individuals, but not completely independent “substances”. When we do something, it is also the universe doing something. It is not a threat to our existence that it is tied to that of the rest of the

universe. On the contrary, that is what gives it meaning and context.

I will get back to this topic later in 9.1.2, but for now, we have to somewhat put it aside and return to those things we can analyse. Do note, though, that the more specifically analysable arguments do bear on this question too. Which way should we understand meaningful authorship? Well, arguments about which conception is otherwise meaningful could certainly guide that.

4.3.3 Compromising on intuitions, preserving what matters

> [T]hrough the dust of three centuries of debate, I think I discern some writing on the wall: if no amount or kind of cognitive or volitional capacity and complexity that could obtain in a deterministic world will suffice for free agency, then simply adding the requirement of indetermination will not suffice either. That means that either free agency is ineffable, free agency (or some significant part of our conception of free agency) is illusory, or compatibilism is true. Take your pick (if you can).
> - Gary Watson[269]

After all this preamble, the argument I will make for my endorsed choice between the agent-causal or compatibilist conception of free will can be stated very simply.

The case for the agent-causal conception is that we have intuitions to the effect that free will requires indeterminism. The case against it is that it means we lose control of ourselves in a concrete sense while calling the loss freedom (and that we have various intuitions to the effect that determinism is required for freedom).

The case for the compatibilist conception is that it allows us to have concrete control and most other things we want from freedom (and that we have various intuitions to the effect that free will requires determinism). The case against it is that we have various intuitions to the opposite effect.

[269]Watson 1987, p. 168.

Unless following your intuitions is the most important thing, more important than being able to control yourself and all other aspects of free will that are not reducible to the requirement of indeterminism, the choice should be clear. We lose nothing in discarding the strong and multifaceted but still singular intuitive requirement of indeterminism. We lose everything else by embracing it. If we simply stop requiring indeterminism of freedom, the conception we are left with has no other problem beyond that (at least as far as everything covered in the above extensive discussion goes). In particular, it gives us everything independently desirable about freedom. If we had indeterministic free will, we would compromise those other things to the extent that the indeterminism has an effect.

The intuitive demand that free will requires indeterminism seems mistaken. There is something seriously wrong with it in the sense that it would only give undesirable and absurd consequences. If we want to have a useful and consistent definition, we should go with the compatibilist one. I can also promise that I can make much use of it later on for things such as explaining responsibility.

Someone could still insist that, under postulated ownership, the agent has control of their actions and is free and responsible in just the right way. I suppose they could also claim that the agent has bodacity and is appletastic. If we are to talk about free will as something that is important and meaningful in our lives, we need to pay attention to what consequences the different conceptions of free will really have for our lives – questions such as whether we would have to fear making the less desired choice for no reason, or whether we could really deliberate meaningfully. It is perhaps understandable that such considerations can seem to point to incompatibilism as well, but that turns out not to be so at all on closer examination. Based on its effects, the deterministic model wins clearly – even when its details have been so poorly filled out at this point.

Based on all this, I could formulate an argument that compatibilism is *really* the more intuitive view, and it might even be right. After all, as pointed out in 4.3.1, it seems as though people want to be able to think there is indeterminism while having the *consequences* of determinism. However, I will stand by what I said in

2.2.1 about the empirical evidence being about equally balanced, besides of which, I should engage with the existing discussion about how to interpret the evidence to begin to make any such strong statements. Thus, I will claim no more than this: the foregoing discussion seems to show compatibilism really has much going for it even intuitively.

4.3.4 Libertarian is not better

The title of this subsection could be, more verbosely: If the libertarian option is self-contradictory, then it is not also better or somehow richer.

Some philosophers think the opposite: libertarian freedom, even though they think it is impossible or self-contradictory, would still be a higher form of freedom than what we can have. At the most extreme, as will be described later (6.2), Saul Smilansky[270] considers compatibilist freedom and responsibility as partial and imperfect, whereas the true version of both would be libertarian – even though that is impossible according to him. In a similar vein, Ted Honderich's notion of two kinds of life-hopes (see 3.4.10, above) suggests it would be better if we could fulfil both kinds[271]. On the opposite extreme from Smilansky, we have Manuel Vargas saying that "it might be nice if we were libertarian agents, but we likely aren't."[272] This barely counts as disagreeing with me, but "it might be nice" is still a slight concession.

I do not think this is the correct thing to say in any version. Once again, one part of the randomness argument was to show that there is no independent reason to desire indeterminism in terms of freedom. Overall, it has been shown that the idea of freedom as indeterministic leads to serious problems, whereas the compatibilist view is essentially unproblematic. As there is nothing left that is worth wanting about

[270]Smilansky 2000.

[271]Honderich 2002, chapter 8.

[272]Vargas 2013, p. 15.

libertarian free will, there remains no reason to say that libertarian free will would be a higher level of freedom. Analysis reveals that it is not. It is a little like saying that a five-angled square would be a better square than the ones with four angles, perhaps adding that it is unfortunate how the definition of a square forces us to live with only these partial, imperfect squares.

That said, I will soon (chapter 5) present the idea that we may *learn* something from the incompatibilist intuitions combined with the compatibilist conclusions reached here to find a view of free will that is *actually* better. They can be *almost* right without being really right.

This is also a disagreement I have with many hard incompatibilists: if they think libertarian freedom fails conceptually,[273] why do they insist on treating it as the only option? The question is half answered and half begged by the idea that getting rid of the notion of free will would itself have positive effects.

4.3.5 The case for (and against) hard incompatibilism or other free-will scepticism as a positive choice

I have been saying that we do not have to choose hard incompatibilism, that we can keep belief in free will via a compatibilist approach instead. It is also possible to take the stance that the belief in free will is undesirable. I have not seen anyone outright express the view that both options are plausible but we should choose the hard incompatibilist view. My main example of a hard incompatibilist, 'Trick Slattery[274], repeatedly makes the case that the incompatibilist view of free will is the real one and compatibilists want to replace it with something else. However, the idea that to be released of the idea of free will is beneficial is also a clear motivation for him, and he might not be so insistent on opposing compatibilism if he did not think it tries

[273]They can even think that is it useless and unnecessary, so making an argument similar to the one I make in this section, but still hanging on to the idea that it is the only real option. See e.g. Slattery 2014, chapter 42.

[274]Slattery 2014.

to bring back pernicious ideas of freedom and responsibility.

Gregg D. Caruso is also worth mentioning here as an example in spite of not being a true hard incompatibilist. He thinks that incompatibilist responsibility is a coherent position, but the best scientific and philosophical arguments disprove it; our acts do not rise in the right kind of way to count as free. As such, he advocates a model of quarantining wrongdoers instead of punishing them. His discussion with Dennett shows how he thinks this will address the wrongs of the current justice system in the United States, but at the same time how Dennett thinks about the same things can be addressed in a compatibilist system of punishment.[275]

It may seem odd to say that we *need* to give up on the ideas of free will and responsibility when this is what is often seen as a threat (see e.g. 2.3, and chapter 6). The fear could be e.g. that giving up on responsibility will only cause chaos where people feel they have a licence to do anything, and our normal relationships will no longer function.[276] This is an intelligible idea, but so is the converse idea of the no-free-will (NFW) advocates: that realising we have no free will is going to make us more understanding and compassionate, especially erasing the scourge of retributivism, of wanting to punish people because they are thought to be responsible, and enable us to act more realistically in the world as it is.[277]

I do think the concept of responsibility plays an essential role, which I describe especially in 9.5, and indeed, this is part of the reason I want to retain it. However, I am highly sceptical about all overarching claims about what the effects of the NFW meme would be on society. I doubt that it itself is going to make us vastly more compassionate or to run amok. Some versions of it given more particular

[275]Dennett & Caruso 2021.

[276]See e.g. Slattery 2014, chapter 34 and Sapolsky 2023, chapter11. Cf. also Smilansky 2000, chapter 7.

[277]See e.g. Harris 2013, pp. 63–65, Slattery 2014, p. 267 (for a single clear example among many in that book).

interpretations might be able to do either.[278]

Understanding and accepting what I have argued above, including that the indeterministic compromise is untenable, effectively forces one to choose either compatibilism or hard incompatibilism. It does not force either or the two choices above over the other, however. I would not be entirely opposed to a hard incompatibilist position that involves understanding what we are actually talking about, such as 'Trick Slattery's view. I do still advocate the compatibilist alternative instead.

One way of putting my reason for preferring the compatibilist alternative is that hard incompatibilism concedes too much to libertarianism. (Naturally, hard incompatibilists can accuse compatibilists of wanting to be libertarian even though they realise they cannot.) Hard incompatibilism based on the randomness argument and other reasoning above is simultaneously proving the absurdity of the libertarian definition of free will – and asserting its correctness as the only possible alternative. This is hardly justified given that people have mixed intuitions, that the compatibilist alternative has existed through history, that it is now the majority opinion of philosophers – and that it is a better normative analysis assuming we care about the concrete implications.[279]

Not only is the compatibilist option presented here a better analysis than the libertarian one in terms of providing what is important, it is also more fruitful. If we understand the matter properly, we will see that the libertarian option is a dead end, and the hard incompatibilist agrees with this. However, with the compatibilist option, we can go on to see what we should say about concepts like the ability to do otherwise and moral responsibility and, as I will demonstrate in the rest of this thesis, "re"-interpret them in a similarly fruitful, logical and actually quite intuitive way.

[278]Cf. Sapolsky 2023, pp. 265–266.

[279]At this point, I am referring to the implications for what we should want of free will, which has already been discussed – not the possible implications about moral responsibility that free-will sceptics might be even more concerned with.

We can do this while avoiding pitfalls of uncontrolled indeterminism, impossible ultimate origination, as well as intrinsic-value retributivism. (See especially chapter 9.)

I see no reason to think (as some advocates seem to[280]) that spreading the NFW meme is going to be a powerful cure for the ills of libertarianism[281] and retributivism, and consequently that compatibilism is a quasi-libertarian threat that will keep people from seeing the light. There is only so much understanding mere labels can convey. Still, it is not that the compatibilist meme is any more likely to be a cure-all. I am choosing a compatibilist approach and terminology here as one option, and I do think it is one that works, arguably even the best one for refuting the mistake of incompatibilism, but not the only one.

4.4 Conclusion to chapter 4

It was noted above (1.1) that the problem of fitting free will and determinism together can be called the "compatibility problem", whereas the corresponding problem with indeterminism is the "intelligibility problem". This chapter together with the previous one has shown that the compatibility problem is indeed, and merely, that determinism *appears* at first sight to be in contradiction with freedom. This is because, meanwhile, the more closer we look at the intelligibility problem, the more clear it becomes it is a real problem. We do not have a possible sensibly intelligible understanding of indeterministic freedom. All we gain from determinism is randomness, and conversely, all we lose with determinism is absolute randomness. Even then, if we actually need randomness, we can still have pseudorandomness,

[280]To some extent, 'Trick Slattery as mentioned above, but even more so collectively the members of his Facebook group "No Free Will: Determinists and Incompatibilists". At least the meme that spreading the meme will help has been successfully spread among this select group.

[281]In this case, perhaps also for the ills of political libertarianism; see the discussion on social welfare and individual responsibility under 9.5.

which is just as good[282].

Following from all this, I adopt a compatibilist view of free will as a compromise that barely compromises anything. As for why it feels as though it does – that is discussed in the following chapter and used as a springboard to understanding more about what is really going on with free will.

[282]See 3.5 above.

5 Beyond Determinism: The Special Nature of Freedom

> But if free will and responsibility are compatible with determinism, then we need to know more about what free will and responsibility are that will explain why the appearance of incompatibility is so persistent.
> – Susan Wolf[283]

> If the conflict between free will and determinism is a confusion, it certainly is not a shallow confusion.
> – Robert Kane[284]

In the previous chapter, I made the argument for compatibilism over incompatibilism: It is common for some intuitions about free will to imply indeterminism and others to imply determinism. Intuitions to the effect that an act is not free if the agent could not have done otherwise, and to the effect that it is not if something else in the past than the agent's free choice was a sufficient condition for the choice happening, can easily be taken to imply freedom requires indeterminism. On the other hand, other intuitions about what it means to have control over one's choices are on closer examination more compatible with the idea that freedom requires determinism. This leads to a position where our "intuitive concept of freedom" is self-contradictory. However, there is a very effective way out of this

[283]Wolf 1990, p. 24.

[284]Kane 2017, p. 2480.

strange and bleak-sounding conclusion. First, we observe that just about everything we implicitly expect of free choice, and that is valuable about it, is compatible with it being a deterministic process. Second, we observe that the intuitions (or strict interpretations of intuitions) that require indeterminism only introduce useless randomness into the process – they do not give anything else of value, instead only take away the real value we can achieve with deterministic freedom. Thus, we could conclude that these (interpretations of) intuitions are somehow faulty or meaningless, and that there is no reason not to just adopt a compatibilistic view of what it is that we should want from freedom that is largely if not completely intuitive as well as otherwise purposeful.

In this chapter, we nevertheless return to the intuitions inspiring incompatiblism and libertarianism and ask just why those intuitions are there in the first place. In one sense, the question is not terribly relevant because the argument for compatibilism has shown that the incompatibilist version of those intuitions does not particularly deserve to be taken seriously as a model of free will. In another sense, however, it makes the question all the more pointed: Why is it that, with so little reason to hold incompatibilist opinions, they still persist, even after compatibilist arguments?

To attempt to fully answer this question would require an extensive collaboration between philosophers and empirical scientists. However, I believe I can shed some light on what part of the answer might be without a proper empirical investigation – at the same time as I look deeper into what the special nature of freedom is like. After all, the argument of the last chapter showed that determinism should be seen as a more or less *necessary* condition of freedom. To claim that it would be *sufficient* would still be absurd. We do not think every bit of the universe is free if we are soft determinists. If we restrict ourselves to considering agents, we also do not find that every agent that might not be considered free in some situation is so because they are indeterministic. On the contrary, someone like an addicted person compelled by their addiction might be "especially deterministic". Thus, the question about (in)compatibilism is just a starting point in my discussion of how

freedom and responsibility should be understood. Affirming compatibilism does not yet tell us what freedom is so much as what it is not. In my quest to uncover the real nature of freedom, I will next look at what it is about freedom that makes it seem as though incompatibilism is true.

I will start by looking at what the "incompatibilist intuitions" are again. Next, I will present two thought experiments that show that *even assuming determinism*, it is understandable that creatures making choices would feel that they have different options open to them, and that the prediction of free choices would be a problematic. After this, I will argue that we seem to think of events and actions under three different modes of thought: causality, randomness and agency. I suggest that a confusion between determinism in nature and the causal mode of thought is what leads to the feeling that determinism contradicts freedom – for the modes of thought all contradict each other and are hard to apply at the same time. The features of the "agency" mode of thought take us closer to the answers of what actual free agency is like. Finally, I look at a requirement for freedom that is part of my own thesis in this dissertation: that while free actions are deterministic considering the total state of the universe at the time of decision, they cannot be describable as determined in terms of coarse-grained, higher-level laws.

5.1 The intuitions behind incompatibilism

Since this chapter revisits, draws from, and seeks to some extent explain the intuitions behind incompatibilism, I will start by summarising the intuitions in question.

A major motivation for incompatibilism and libertarianism seems to be the intuition that, in a free action, we are able to do otherwise. We saw that when this is taken to mean indeterminism, it does not work as a requirement due to its unsavoury consequences. In section 5.2.1, we will look at how it might be unsurprising that we would have such a feeling even if our decision-making *were* deterministic. In section 5.4, we will see how there is a reason, even in a compatibilist view of freedom, to formulate a requirement of doing otherwise that contradicts certain kinds of

determinism.

Another major intuition that can be interpreted in a way that motivates incompatibilism and libertarianism was that we should be the origin of our own choices. This could be interpreted as referring to each choice being uncaused in the sense of being caused only by some vague will or self with no properties that fix the decision, or to our having made such a self-originating choice earlier, leading to our character now and our choices based on it. In section 5.3 below, I will show how this lack of further origins in both forms – in the individual choice event, or in the character behind choices – is part of a mode of thought that I propose we usually use when thinking about intentional action.

Prediction is closely associated with determinism, and if predictability were somehow contradictory with free will, understanding why might give some perspective as to why determinism seems to be so too. The second thought experiment in section 5.2.2 shows how this is true in a sense even assuming determinism, and in section 5.4, we will see further why it is important for real freedom.

I also suggested in 2.2.2 above that incompatibilists seem to be working on an intuition that there is determinism, there is randomness, and there is some third thing, an indeterminist freedom that is not randomness. I argued that this is impossible and a confusion (3.2.1). One of the main points of this chapter is to answer the question of just what they might be thinking of when they think of freedom as a "third thing". This is answered in section 5.3, and I think it is the most important part in this chapter for understanding the intuitions and thought patterns leading to incompatibilism and libertarianism, though it is only a stepping stone on the path to seeing how freedom should really be understood.

5.2 Two thought experiments concerning determinism and freedom

In this section, I present two thought experiments that assume determinism behind our choices but come to conclusions that contradict what is assumed of determinism.

I call these thought experiments *the decision machine* and *the prediction machine*.

5.2.1 The decision machine (and the feeling of being able to do otherwise)

First, suppose that a deterministic universe contains creatures that are sentient and sapient – so that they can feel and think as if things are one way or another, to have intuitions and interpret them – yet are as deterministic "machines" as everything else in the world. These creatures make decisions or "decisions" by taking in all the input of what they know about their surroundings, comparing it with their goals, and determining the "best" course of action according to a deterministic process that is always capable of reaching only one unique outcome in each unique situation. Or perhaps they are not really so rational as this makes it sound. Perhaps they are determined at each moment also by various whims or unconscious motives, like we are, and do not arrive at the "best" decisions, yet it is still true that their decision-making process can only lead to a unique single outcome in each unique situation. That makes little difference for the outcome of this thought experiment, as long as we make sure to say that, like ourselves, the creature is at least sometimes conscious of being in the process of deciding something.

The question is: based on what has been said, what could such a creature or decision machine be expected to feel or think at the moment that it is making a decision?

The decision-making is a deterministic process that takes some amount of time. It goes on inside the creature, involving parts that are parts of it; causal chains lead to it from elsewhere, but at the time this process is underway, it is happening inside the creature. Nothing in this scenario allows the creature to certainly predict the result of the process in advance, either, just to go through it and then see what the result is.

The creature is (let us suppose so in case it is not necessarily true) also like us in that it thinks of multiple options that it weighs up against each other, and it can potentially keep coming up with new ones to add to the list. Making the decision is

thus a process of coming up with options and considering which one of them to choose.

What this all leads to is that the creature will, while the process of deciding is underway, *have several options in front of it between which no determination has yet been concretely made*. Further, its mind is what is working on making the determination, so the different options are in its mind, being treated as options while the process of determining just one of them as a choice is underway. Further, the creature is at least sometimes aware of this.

If what happens with us is something like this, determinism and all, it is not surprising that we have the feeling that different options are open to us. They are, just not in the useless sense of indeterminism. Even if you want to say they are not, the *feeling* is still unsurprising, as we are processing the options.[285]

If the creatures did not feel that way, how could they decide anything? They may feel they will only choose one option (which is true under indeterminism too: see 3.4.11), but they cannot think they have to choose one particular, identified option before they have gone through the process determining which one it is that they will choose. Only the finished process of deciding determines what the result will be.[286]

This is neither proof that we would *have* to feel this way under determinism, much less proof that our feeling is proof of determinism. The argument is not strong enough and takes no account of possible interfering factors.[287] What it does show is

[285]Cf. Dennett 2015, p. 123.

[286]Cf. Donagan 1987, p. 160: "Yet there is a sense in which you cannot plan a plan: namely, that you cannot deliberate about what a specific result of deliberation is; for if you can specify it you have nothing to deliberate about." Similarly cf. Dennett 2015, pp. 122–123.

[287]Not strong enough, that is, to prove the first claim that we would have to feel so. Deriving from the argument the second claim that our feeling would prove determinism is of course just logically fallacious. I only took the time to dismiss it to be as clear as possible.

that it *makes perfect sense* for us to have an intuition about alternative possibilities if determinism holds of our decisions.

I describe an argument in 2.5 that says we cannot make (conscious) decisions if we perceive ourselves as unfreely determined. This is another way of saying why. Before we make a decision, we can at most think that we *will* be determined some way or another; if we say we will be determined to some specified outcome, we are saying we have already chosen it, though we might still change our minds after that. Again, as Hartshorne put it when trying to argue for the other side:

> *Before* I decide I may claim to know that my decision will be fully determined, whether by heredity or environment, or by God, but in what way can my decision take this alleged knowledge into account? *After* the decision I can say, See what I was preprogrammed to decide![288]

This gives us some vague idea about two different things: what might be the cause of the commonness of the intuition that freedom requires the possibility of doing otherwise, and what the role and interpretation of being able to do otherwise might be in a better theory of freedom. If we consider multiple options, it makes sense for us to feel that they are still open while we are considering them; and if we *cannot* consider multiple options at all for some reason, we are not very free to make a choice. The idea of open options was already discussed in 3.4.9, and I will return to this point from a different angle in 5.4 below.

5.2.2 The prediction machine (and the non-predictability of free choice)

The following thought experiment aims to demonstrate that, even though determinism in some broad sense implies predictability[289], even deterministic agents

[288]Hartshorne 1984, p. 19.

[289]See 2.1.1 above.

can become impossible to predict in every way under certain circumstances – namely, when they can interact with the prediction.[290] Understanding why helps us gain more perspective on how determinism and agency can be seen as harmonious, and it also demonstrates something about what is important for freedom.

After formulating this argument by myself, I became aware that the same basic point was presented under the name "the paradox of predictability" by Michael Scriven over half a century before.[291] Thus, the prediction machine thought experiment as I present it can be considered as a version of that paradox, though I do not regard its results as a paradox that needs to be solved, merely as something surprising but understandable.

Imagine a computer having all the information and processing capacity of the Laplacean demon (see 2.1.1), so that it could predict with certainty what will happen in the universe at a given time – including the actions of human beings. If we suppose that it was separate from the universe in the sense of having no causal influence on it, the picture is simple. If an observer *outside* the universe read the prediction and then observed the universe (likewise without affecting it), it would simply see the universe repeat what the computer had already said.

Now suppose that the output of the computer would be observed by an actor within the universe itself. To predict that observer's behaviour, the computer would have to take into account the causal influence of the output itself. Suppose that without the computer having a causal influence on the universe, the agent would have taken actions that she would not have realised would cause consequences highly undesirable to her. Then, if the agent were to see the output, other things being

---

[290]Cf. Dennett 2015, p. 11; p. 123, footnote 12.

[291]See e.g. Gijsbers 2023. This paper also makes a number of points parallel to mine, though, again, I formulated mine before reading it. The original idea is credited to a 1965 paper by Scriven that I was not able to find. A similar thought experiment is also formulated in Slattery 2014, pp. 227–228. Footnotes later in this subsection in the present work will show that many others have discussed similar ideas.

equal and if she were sufficiently rational, she would change her mind and act differently in order to avoid those consequences, falsifying the prediction.

To make things more difficult, suppose an agent who feels his free will is threatened by the machine's predictions and who would thus see it as an undesirable consequence that the prediction, whatever it may be, would be correct. Such an agent would, insofar as he was able to, seek to always act differently than predicted. Note that this supposed freedom is entirely in harmony with total causal determination. The choice of acting "otherwise" would always causally follow from the agent's state and the output of the computer.[292] This would just be a case of the agent acting based on information received from the world to effect a desired outcome.

Given that the computer would aim at making an accurate prediction, and would do so successfully, what would happen in such cases? It seems that the answer is so far indeterminate, requiring first a more precise description of the general program that the computer would run. In fact, not taking this kind of scenario into account when designing the program could lead to the program not completing its task successfully.

The computer could perhaps give a conditional answer: This would have happened, but if I had predicted it, this would happen instead. This would not be enough, though, at least in the case of the agent unwilling to be predicted, since he could still falsify the prediction. Giving this compound prediction would also change what happens afterwards yet again, so that a third clause would have to be added, and this could change the consequences yet again. The conjunction might be infinite; to avoid this, the computer might have to somehow calculate how it might reach equilibrium.[293]

---

[292]Cf. 3.4.10.

[293]Notice that what is happening here is that the description of a computer that would perfectly predict what would happen has been shown to be ambiguous; it described a goal, but now we would need to know what it does in order to achieve that goal, leading to the question of whether it would even be possible. Ironically and/or fittingly, this is part of the same problem described for the reasons-responsive agent potentially following higher-level laws below (5.4) and starting in the present section.

Reaching equilibrium would mean giving an output that causes itself to be correct. In the case of the agent who merely wished to avoid some negative outcome, the output might be a prediction that the agent would see as good advice, leading to a desired outcome, thus causing (motivating) her to act exactly as predicted. In the case of the agent determined not to be determined, the prediction would have to cause the agent to act otherwise than he had intended. The prediction could be accurate but too confusing for the agent to be able to decipher it, too complicated perhaps, causing him to unknowingly act as predicted. It could also be persuasive, changing the agent's motives, perhaps in the form of a prediction that the agent will accept a new compatibilist view about freedom, describing the view in such a way that the agent would indeed be persuaded to accept it.

Suppose that the agent was perfectly "rational" in following his goal to be unpredictable. He would have godlike intellect that would make him impossible to fool[294], and he would be immune to persuasion. Whatever the computer predicted he would do, he would do otherwise. Then, *while still assuming determinism,* it would be impossible for the computer to predict his actions other than counterfactually. The computer could not make a single exact prediction that would be true of the world, in spite of its Laplacean intellect and knowledge. It could still conditionally know everything about the future of the world and the agent's actions.[295]

In a sense, this is a paradox (but not contradiction) related to considered, intentional action and of freedom. In a different sense, it is not, because you will get the same kind of results without assuming that any agents are involved. If you suppose that there are no conscious actors in the universe, if the computer's output appears in the universe – say in the form of a big screen floating in space with no-one looking at it and displaying a part of the results – it still has a causal influence

[294]Putting aside the question whether such a mind could exist within the universe.

[295]Cf. David Chalmers's concept of conditional scrutability (2012, pp. 53–58). Scrutability is briefly explained in this work in section 10.1.

on what happens, and will have to take that into account in its predictions, again necessitating its either giving conditional predictions or looking for an equilibrium in which its output partly causes itself to be true. Given complicated enough reacting systems, this can also lead to the impossibility of prediction again. The screen could, implausibly but logically possibly, affect the universe in such a way that it would always make its own predictions untrue. Any system that effectively acts so as to avoid the prediction being true is similar to an agent in this limited sense. This suggests a perspective for looking at freedom: our agency is not an acausal mystery but a sophisticated application of causality – not apart from (the rest of) nature but on the same continuum with it all. (See 9.1.2.)

If it comes to that, we can also question the freedom of our earlier hypothetical agent if he always invariably acts so as to invalidate the prediction.[296] Imagine a prediction that, in part, says "Part of this prediction has the causal effect of causing a war that will kill the agent's family unless he does exactly as this prediction says." While there is no absolute logic saying this prediction would have to come true, a free agent would typically give in in the face of such a threat, specifically because he *is* capable of changing his mind to seek the better consequences rather than being determined only by any single motive or causal thread. Of course, acting under a threat is often not considered free action. I do not dispute that. In a sense, the agent would not be free on this occasion. However, the agent would likely be persuaded by the threat if he had the qualities describing a free agent generally, qualities that are necessary rather than sufficient conditions for being free. To be persuadable for rational reasons is implied by freedom as usually understood.

If you have a system where part of the system is making predictions about itself and feeding them back into the system, you get a cybernetic, self-referential feedback loop.[297] In the case of a computer making a prediction in advance and

[296]Cf. Fischer & Ravizza 2000, p. 45.

[297]Cf. Dennett 2015, pp. 38, 70–71 for similar points.

seeking to give an output that will be correct even given their own causal influence, the loop will take place *within* the computer. In any case, such systems, even though they do not contradict causality, become difficult to describe under it.

Thus, we can conclude that freedom seems to contradict predictability in a limited but strong sense. In particular, an agent that would see a prediction of a future with undesirable but avoidable features would not be very free if they could not go against the prediction. At the same time, we saw that an agent that would be absolutely adamant in following a single motive would seem to be irrational in such a way as to seem less than free. The reason both of these examples contradict freedom lies, I will argue, in the requirement of reasons-responsiveness, further discussed in 5.4.1 and later in 8.1.

For now, we can conclude from the previous that we have found a possible explanation for why it seems as though predictability contradicts freedom: in some sense, it does, even with compatibilist assumptions. The mistake is equating predictability of the relevant sort with determinism.[298, 299]

---

[298]Some related references:

Denyer 1981, p. 92 (italics in original) presents a similar point as a refutation of determinism in human behaviour: "[Scientistic determinists] acknowledge [the point about human unpredictability] by saying that if a prediction based on a conjectural law of human behaviours is communicated to its subject and if its subject is encouraged to falsify it, then of course one *interferes* with things so as to falsify the prediction and the conjectural law on which it was based." It should be clear that I have just answered this as an objection to determinism pertaining to humans. However, the point that there are no laws of human behaviour because of this will be taken up again in 5.4 below and developed further from a compatibilist perspective.

See also Honderich 2002, pp. 83–86 for another thought experiment that brings confusing results which can, I think, be explained by the logic of the prediction machine. Note also the idea of "stepping-back" (*op. cit.*, pp. 86–87).

Other similar points related to the recursive unpredictability of agents are made in Almond 1998, p. 18; Dennett 2015, p. 42; Ofstad 1967, pp. 183–184.

[299]This problem relates to self-referentiality and, for that reason, is structurally similar to the liar paradox. Many paradoxes in general follow from similar self-referential recursiveness: see Yanofsky 2019.

5.3 Three modes of thinking

In this section, I argue that the discussion about freedom and determinism can be described in terms of three kinds of ways of thinking about why things happen. These three can be shortly described as causal explanation, appeal to randomness, and appeal to agency. Further, I argue that these ways of thinking may have a strong, pre-existing psychological existence, because that would explain the debate so well.

5.3.1 The tension between agency and causal explanation

It seems that our sense of the agency of actions tends to disappear when we look at things from a deterministic perspective – or, more precisely, from a perspective that causally explains the actions as events. If the agent was caused to make their choice by events outside of themselves, then it suddenly seems as though the agent is not making a difference, or is not in control.

I argued in 4.3 that we need not say this: the agent can be in control in the only meaningful way available under such a deterministic scheme, and we should not confuse the idea of being the *origin* of one's actions with the impossible notion of being their *ultimate* origin.

For all that I know this, I can still "make agency vanish" by thinking about chains of causality. (See also 1.1.3 and 2.5.) I can make arise in myself the feeling that agency is not present in the picture by thinking about the prior causal chains. We can ask: how can an agent be free if their choices are pre-determined by something they had no control over – by conditions before their birth? We can also ask: how can anyone be responsible for their actions if their actions are thus determined? Personally, after all the exposure to compatibilist thinking, I find this latter perspective even more compelling than the one about loss of freedom.

In the meantime, I will call attention to something that may be missed. We can ask "How can we be free/responsible when our choices are determined by something outside ourselves?" and it will probably sound perfectly reasonable to most people. Yet, we are not actually too clear on *why* it seems reasonable. Clearly, there are some

unsaid premisses at work. It has something to do with the concept of agency, obviously, but the arguments in chapter 4 showed in some sense that free will and agency are not threatened by determinism. Nevertheless, that does not keep us from asking the question, and it may even sound as though it makes intuitive sense, even to a compatibilist such as myself.

If we have such *feelings* in spite of the best *arguments*, perhaps it is better to look into explaining the feelings *as* feelings.

5.3.2 Three modes: randomness, causality and agency

I will be modest in my first claim in this subsection. Based on the previous discussion of what people think about questions of freedom, responsibility, and determinism, we can say that one correct way to describe these ideas is that the following ways of thinking (and feeling) about how things happen appear in the discussion.[300]

I will mention the simplest one first. People think about things happening randomly, meaning there is no explaining why something happened and no way of predicting what will happen. Everything I said previously about what indeterminism really means probably applies when things are being thought of this way.

Secondly, and still relatively simply, people can think of things in a way that involves their being causally determined under (more or less) deterministic general laws. In this case, an event is explained by referring to previous events and laws. Additionally, in this way of thinking (not in all causal thinking but in what I am defining here and claiming to be one way of thinking found in the discussion), nobody is seen as being free or responsible. This last apparently has something to do with how their actions are explained by factors outside them. When we take this stance with the actions of agents, we are usually either excusing the agent as not responsible or talking about the paradoxes of freedom and determinism and getting confused.

[300]Cf. Dennett 2015, pp. 120–121.

Third, and most complicated: This way of thinking is a way of thinking about an agent's actions, and it has the following properties:

- The agent is seen as being free and responsible.
- The agent is seen as being the origin of their own actions.
- As shown by the compatibilist arguments in chapter 4, the agent's actions are implicitly seen more as if they are deterministic than as if they are indeterministic. However, the question of which they are is not, I think, properly answered even implicitly.
- People taking this stance often, but not necessarily, assume that what it describes contradicts determinism, but also contradicts randomness.
- This way of thinking somehow depends on stopping the analysis of the chain of causality leading to an action – stopping it on an agent's choice or property (i.e. character trait).[301, 302] It has been shown above that, for many people including myself, if we start to go further back in the analysis, we naturally find ourselves thinking in the previous, causal mode instead, losing the sense of agency, freedom and responsibility.
- This way of thinking also seems to *encourage* stopping the analysis within the agent. People seem to easily fall into the trap of thinking it makes sense to say something is a choice instead of being either determined or random, even though that makes no sense on closer analysis.[303]

Based on all this, it seems reasonable to propose the following hypothesis: the

[301]Cf. Denyer 1981, p. 66: "One is then committed to the belief that there will be future events, one's own future actions, whose origins will lead back to one's own deliberations but no further[.]"

[302]There are differences between it being about a character trait or about a choice, and my having to lump these options together represents one way in which my model is incomplete. However, it seems to also reflect a tension within the way people actually think about this.

[303]Cf. Honderich 1973, p. 210.

three modes of thinking proposed have some kind of psychological reality to most or many people, and that is what is causing much of the confusion in the determinism–free will debate.

To reiterate: Above, I claimed that what I have already said is enough to establish that it is reasonable to say that the debate can be seen as manifesting all of these three modes of thought as I defined them here. The further claim that I am making is that it is a reasonable psychological hypothesis that many or most people (out of those groups whose opinions have been surveyed here, at any rate) have a significant tendency of some sort, possibly biological, to think in modes that are something like the three ways of thinking I described above.

The reason to think this is plausible is that it explains the debate when it has already been established that some of the views expressed are not reasonable as such. It explains why the convinced compatibilist still feels as though agency disappears when thinking about causes further back: their way of thinking is slipping from agency to causality. It explains why people tend to feel determinism contradicts freedom: determinism evokes the mode of causality, which deactivates the mode of agency and at any rate works differently from it. The idea of modes of thinking also explains why both determinism and indeterminism seem to contradict freedom: the modes of causality and randomness both contradict the mode of agency, and the mode of agency involves stopping explanations within the agent, not affirming either determinism or indeterminism beyond the stopping point. It explains why it is easy to feel satisfied with the vacuous explanation that an agent causes something: again, the stopping point. Finally, it explains the persistence of incompatibilism even given what we have already seen proven about how well compatibilism works.

It makes perfect sense for such categories to have evolved as part of our thinking, whether biologically or culturally. Sometimes, there is no meaningful reason to be stated for why something happened, hence randomness. Sometimes, we can see, explain and predict the universe acting lawfully, hence the causal mode. Sometimes, we need to be aware of who did what and hold them accountable, hence agency. Sometimes, the only reasonable thing to do is to excuse someone of

responsibility because of circumstances they could not help, hence the possibility of applying the causal mode (or randomness) to an agent, and the associated feeling that responsibility vanishes.

Again, I am making my claim as modest as possible by saying that it is a possible explanation that some such pre-existing modes of thinking are the explanation for intuitions about determinism and freedom that have come up in this discussion. This also implies I cannot say very precisely what they would be like as common psychological phenomena. That said, I will nevertheless define and characterise the modes of thought as I envision them more precisely, so that I may refer to something more precise when referring to them.[304]

All of these are ways of thinking about an event and its causes – about how it can be explained and what follows from that. They also evoke certain kind of feelings, such as anger for a wrong committed by an agent in the mode of agency or feelings of whether something is an instance of freedom or not.

The mode of randomness has the following characteristics:

- Amounts to saying that something happened by chance or for no reason or no known, knowable or relevant reason(s).
- Events explained under this mode are not attributed to an agent and evoke no feelings of freedom, agency, responsibility, wanting to praise or blame, etc.
- If applied to an agent, normally implies some kind of insanity.
- Explaining events under this mode instead of attributing it to an agent in the mode of agency can be used as an excuse, to evoke the feeling that the agent was *not* responsible.
- Corresponds closely to indeterminism, i.e. explaining an event this way

[304]Note that these characteristics are based on my reconstruction of what seems to be going on in the discussion and in intuitions – not pre-existing definitions of randomness or causality, say – and are of course preliminary in any case.

posits it to have basically the same properties as I have argued indeterministic events have.

The mode of causality has the following characteristics:

- Amounts to saying that the event happened for a mechanistic, lawful, natural reason.
- Explains an event by stating that it was caused by another event in conjunction with something resembling a law of nature.
- Events explained under this mode are not attributed to an agent and evoke no feelings of freedom, agency, responsibility, wanting to praise or blame, etc.
- If applied to an agent, normally implies compulsion.
- Explaining events under this mode instead of attributing it to an agent in the mode of agency can be used as an excuse, to evoke the feeling that the agent was *not* responsible.
- Roughly corresponds to a simplified determinism, i.e. explaining an event this way posits it to have properties that are deterministic or probabilistic and close to determinism, but this mode is poorly suited for describing complex deterministic systems with interacting parts, such as those involved in compatibilist free will.

The mode of agency has the following characteristics:

- Explains an event by saying that it was the freely chosen action of an agent.
- May use intentional explanations, deterministic explanations, and sometimes indeterministic explanations, but must involve a part of the explanation that is internal to the agent and is not explained by referring back to reasons outside of the agent.
- Events explained under this mode are attributed to an agent and do evoke feelings of freedom, agency, responsibility, wanting to praise or blame, etc.

- Chaining explanations further back from the part seen as internal to the agent, to the point that nothing inside the agent is left unexplained by things outside of the agent, activates the mode of causality instead.[305] Positing the "internal" component to be random activates the mode of randomness instead.
- Roughly corresponds to a specific kind of determinism or probabilism, in that the mode usually involves explanations of the intentional sort, where reasons determine what the agent does. Sometimes works like indeterminism in that an arbitrary choice or a choice between conflicting motives can be seen as someone's choice while being seen as indeterministic.

These more specific versions continue to explain the oddities of the free will discussion in the manner described above. Of course, the more specific my descriptions get, the less likely they are to be completely correct in a sense that psychological research could verify the existence of just such modules or the like. Again, though, it is quite possible that *something like this* exists as part of the mind's functioning – and explains many of the "incompatibilist intuitions" and their persistence in spite of the better arguments leading to the side of compatibilism.

5.3.3 Two more fundamental perspectives: Mechanism and agency

The three modes above are presented as psychological behavioural modes. They can help model reality reasonably, but they can also be misleading. In light of this, I want to look at something slightly different that they should not be confused with. I will briefly explain two true perspectives on reality related to the same questions.

The first perspective is the *mechanistic perspective* on reality.[306] In this

[305]Cf. Dennett 2015, pp. 36–38, Dennett 1973, pp. 170, 173.

[306]In spite of the name I choose for it, this perspective does not necessarily correspond to

perspective, we only look at what caused what or what events were uncaused. We do not look for agency or make moral judgements. They are not part of the vocabulary or syntax we are using. That does not make such judgements wrong in any way; it simply means we cannot talk about them without changing perspectives. This is a valid way of describing the world and can be used to state meaningful truths.

The second perspective is the *agency perspective* on reality. Here, we do make moral judgements and attribute agency. This is also a valid way of describing the world and can be used to state meaningful truths.[307]

What is important to notice is that, completely aside from anything that I said about the three modes of thought before, if we switch to the mechanistic perspective, we do not see agency-related concepts at all. This is why I think these two perspectives logically underlie the three modes described above; taking either the mode of randomness or the mode of causality leads to one taking the mechanistic perspective, and obviously the same way for the mode of agency and the agency perspective. This is just a contingent part of how the modes of thought as I envision them work, but it is only possible because these perspectives exist in the first place.

The choice to speak of two perspectives here is arbitrary in that I could have spoken (among other things) of three perspectives, each corresponding to one of the three modes of thinking. This is because the choice is a choice between levels of description (or levels of abstraction), as described in more detail in 10.3: such a level describes objective reality, but the manner in which it does is a matter of choice.

I introduce these perspectives here briefly so that, later on, I may refer to "objectively valid" perspectives instead of modes of thought with all their baggage and psychological and hypothetical nature. I will refer to them a little in 5.3.4 right below and later in 6.3.3.

---

what is historically known as "mechanism".

[307] On how moral truths can be meaningful statements about a universe that can also be described from a mechanistic perspective, see Kokko 2018.

5.3.4 Similar ideas by other writers

The general idea that there are two or more different ways of thinking, one concerned with something like freedom or intentionality and another or others concerned with mechanical or random happenings, has been proposed in different forms numerous times before. These proposals vary in how close they are to mine, but all could be counted as predecessors to mine, though I came up with my own idea before hearing about most of these. In this section, I briefly summarise a sampling of such ideas, leaving still others out.

A simple version of the idea that there are different perspectives that make it difficult to see how free will could be compatible with things like determinism and analysis is the one proposed by R. E. Hobart[308] with the example of the villagers who thought there must be a horse inside the steam engine (see 10.7): that it is difficult to shift from thinking of a familiar whole with its implicitly understood rules to thinking of how that whole could be made up of "parts".

P. F. Strawson's idea of interpersonal attitudes and the opposed objective attitude posits a difference that is much like the difference between the mode of agency and the two other modes combined:

> What I want to contrast is the attitude (or range of attitudes) of involvement or participation in a human relationship, on the one hand, and what might be called the objective attitude (or range of attitudes) to another human being, on the other. Even in the same situation, I must add, they are not altogether exclusive of each other; but they are, profoundly, opposed to each other. To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided, though this gerundive is not peculiar to cases of objectivity of attitude. The objective attitude may be emotionally toned in many

[308]Hobart 1934, p. 3.

> ways, but not in all ways: it may include repulsion or fear, it may include pity or even love, though not all kinds of love. But it cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships; it cannot include resentment, gratitude, forgiveness, anger, or the sort of love which two adults can sometimes be said to feel reciprocally, for each other. If your attitude towards someone is wholly objective, then though you may fight him, you cannot quarrel with him, and though you may talk to him, even negotiate with him, you cannot reason with him. You can at most pretend to quarrel, or to reason, with him.[309]

Immanuel Kant famously thought that there is a contradiction between our understanding things as happening in a causal order where everything is caused by previous effects and our understanding ourselves as being the absolute originators of our own actions.[310] This difference was apparently explained by his transcendental idealism as concerning the world as we have to see it, not as it may be in itself. I dare not try to engage in Kant exegetics enough to try to expound or interpret this idea in any detail, but insofar as Kant really thought that the apparent contradiction is unsolvable from our point of view and that it can still be explained by appealing to the transcendental, I disagree and take a more boldly moderate position: there is no contradiction we cannot unravel, and if there was, an answer involving a contradiction to this extent would not be acceptable.

Somewhat reminiscent of Kant (perhaps), Colin McGinn (referring back to Noam Chomsky and later quoted approvingly by van Inwagen who also speculates on the same idea[311]) speculates that since freedom seems to be compatible with neither determinism nor indeterminism – though his guess is that it may be with

[309] Strawson, n.d.

[310] Kant 2013, A 444/B 472, A 445/B 473, A451/B479; Koistinen 2008, pp. 127–135; Allison 2012, chapter 9.

[311] Van Inwagen 2002, p. 194.

determinism on the ultimate level and indeterminism on a higher level – perhaps human beings are somehow "hardwired" in such a way that it is impossible for us to unravel the mystery of metaphysical freedom. The solution could be simple and non-philosophical if we had the capacity to understand it, but, possibly, we do not.[312] Compared to such ideas, I present a related but more simple and optimistic solution: perhaps we are hardwired in such a way as to make it difficult for us to grasp the truth of this matter, but if we step outside our intuitions and ways of thinking instead of insisting on applying them, we seem to be able to both move beyond them and to see how they work.

Saul Smilansky[313] speaks of a *fundamental dualism*, meaning that there is an "ultimate level" (henceforth referred to as *Smilansky's ultimate level*, and not to be confused with what I have referred to above as "ultimate level") on which no-one is free or responsible, but there is also a control-compatibilist level on which people can be seen to be responsible. Smilansky's levels are better compared with my perspectives than my modes. Smilansky's ultimate level roughly corresponds to my mechanistic perspective, and his control compatibilist level to my agency perspective. However, there are two main differences between this view and mine. Firstly, Smilansky's ultimate level is supposed to be the truer one, whereas my mechanistic perspective is not. Secondly, and relatedly, Smilansky's ultimate level contains descriptions about whether freedom and responsibility exist or not, whereas my mechanistic perspective does not. In this particular respect, Smilansky's ultimate level resembles my mode of causality and/or randomness with their exclusion of freedom and responsibility more than it does my mechanistic perspective. Smilansky's view will be discussed in more detail in the next chapter.

Perhaps the closest equivalent I have seen to my idea about the three modes of thinking comes from amateur philosopher Mike O'Neill, who has hypothesised

[312]McGinn 1994, chapter 5.

[313]Smilansky 2000, chapter 6.

that the question of free will and determinism is confusing due to its invoking disparate networks of some sort in the mind:

> The situation is this: every time we work within the notion of free will (or anything like it) things can proceed smoothly, but as soon as the notion of determinism comes to mind the network of thought that produces and houses determinism FIGHTS WITH the network of thought that produces and houses free will!
>
> This kind of mini cognitive dissonance happens EVERY TIME the two thoughts come into close working contact, because of the way the two participating networks are opposite in nature and actually oppose each other. It is impossible to avoid this conflict, once both networks come into close enough contact.
>
> The resulting confusion and emotional discomfort from this mini cognitive dissonance INCREASES upon repeated meditation of the riddle of free will versus determinism. This makes the thinker eventually give up in despair that free will can ever be reconciled with determinism.[314]

The idea can also be found outside philosophy. As an example, in the field of futures studies, the *General Frame of Consistency* (*Yleinen konsistenssikehikko*) developed by Osmo Kuusi makes a distinction between not-learning beings, which can be predicted in advance, and actors or genuine learning beings. It is also noted that there is a phenomenon of dual description whereby actors could also be described as consisting of not-learning things.[315] Though this dual description is the thing most obviously resembling the distinction between personal and impersonal modes of thinking or ways of description, the whole notion of some things in the world being described in one way and others in a different way based on agency is

[314]Mike O'Neill, personal correspondence.

[315]Kuusi & Virmajoki 2022, pp. 33–38. I am for simplicity ignoring the category of not-genuine learning beings, and I am generally describing the theory very hastily in interests of space.

also a common point between this theory and mine. In addition, the notion of learning beings that are able to adapt and are not described by simple laws foreshadows the topic to be discussed here next, which concerns what kinds of determinism actually do contradict freedom.

All these examples suggest that the idea of different perspectives for human action and something else is a powerful and usable one – certainly one that has seen a lot of use. Still, only some of the philosophers that have invoked it have used it to truly dissolve and explain the paradox of freedom and determinism the way I do here.

5.4 Freedom and higher-level laws

I argue in 11.8 that what universal causal determinism means for free choices, as for all unique events, is that each individual event or choice is exactly determined given all the facts about the universe at that moment that can have a causal effect on it, and this is different from the possibility that it would be determined given some more limited description of the world. In this section, I will make use of this insight to mark out one difference between a kind of determinism that threatens freedom and a kind that does not.

5.4.1 Reasons-responsiveness as the ability to do otherwise

Reasons-responsiveness has been presented by John Martin Fischer and Mark Ravizza as a criterion for moral responsibility without implying freedom of the will.[316] This means that the agent is able to make their choice based on good reasons they have to do so, as opposed to being unable to take them into account. (We will return to the details as presented by Fischer and Ravizza later (8.1.1)). If fits perfectly well with compatibilism, and here, I want to propose it as a criterion for freedom as well.

[316]Fischer & Ravizza 2000.

Reasons-responsiveness came up in the prediction machine thought experiment in two ways. Firstly, it was shown that free agents could use a prediction concerning themselves as a basis for choosing to act differently. Thus, being free would imply being able to take relevant reasons into account, being responsive to them. Secondly, it was pointed out that an agent that would be absolutely unyielding in falsifying such a prediction regardless of all other consequences would not be all that free. Thus, being unable to take other reasons into account, being unresponsive to them, would imply a lack of freedom.

In both cases, being responsive to reasons would mean being responsive to new reasons in spite of some earlier prediction that one would do otherwise. In the second example, *not* being reasons-responsive would specifically mean that the agent would not respond to new reasons because they would always follow the more general rule of disobeying the prediction they see.

This suggests an interpretation of ”not being able to do otherwise” as opposed to freedom that is based on reasons-responsiveness and compatible with compatibilism: an agent *A* is not free if *A* is not able to do otherwise even given what is, in *A's* view or based on *A's* values and goals, *a good reason to do otherwise*. Determinism in general does not imply unfreedom by this criterion, but it does if it is the kind of determinism that states the agent will always do the same thing given some circumstances, regardless of other circumstances. This harmful kind of determinism can also be expressed by saying that the agent will do the same thing if the circumstances are the same in terms of some *higher-level description*.

5.4.2 Higher-level laws limit reasons-responsiveness

The concept of lower- and higher-level laws is introduced in detail in Appendix B, and here, we apply it to an aspect of freedom and reasons-responsiveness. Consider some of the examples of people acting compulsively: one suffering from exaggerated kleptomania and another one from exaggerated alcoholism. Let us pretend to start with that true determinism applies in these cases on some level of description, so that the person with kleptomania, for example, will *always* steal when they are in a shop.

Then "Person *K* goes to a shop" is a higher-level description (whose lower-level counterparts are of the form "Person *K* goes to a shop, *and*" something). A related *higher-level law* is that "When *K* goes to a shop, *K* steals something." If this law is true, then *K* steals something every time *K* goes to a shop[317] – regardless of whether, say, *K* thinks he will get caught, *K* believes that stealing is wrong, etc. Those further conditions would be part of a lower-level description.

We could make a similar description of the person with exaggerated alcoholism: she will always succumb to the temptation to drink (although certainly defining what is a temptation to drink will be more complicated), even if, say, she would be better off not doing so before going to work because of all the consequences.

These two examples are paradigms of what would (perhaps) not be considered free choice. The agents lack reasons-responsiveness in these cases and have no genuine choice about whether to indulge their compulsion. In a highly caricatured form, it can be said that this is because they are bound by higher-level deterministic laws.[318]

Next, let us drop the pretence that truly deterministic laws apply in real life. Many things can stop a someone with kleptomania from stealing or with alcoholism from drinking. I am not aware of anyone saying that, to truly have kleptomania, one must be compelled to steal every single time. Almost anyone could be stopped from doing almost anything by threatening to kill them and, less dramatically, someone with kleptomania might only give into the temptation now and then but still be under

[317]I take this to be understood to mean that *K* steals at least one thing on every instance between *K's* entering a shop and exiting the same shop. The definition would have to be precise to be part of a law but, then again, I do not think this level of detail matters much in the case of such an example.

[318]John R. Searle (2007, pp. 51–52) makes the point that descriptions of persons doing things because of their desires, or for reasons, do not seem to imply causal determination by those reasons/desires. Searle explains this from a different angle (pp. 52–57), but it can be explained by referring to my point here as well.

a compulsion to steal. Therefore, the rule that defines compulsive behaviour as unfree cannot require actual determinism on the higher level of description. What does it require, then? The answer is, I think, that it requires something *close enough*. It requires a strong enough compulsion to act in a certain way in certain kinds of circumstances that we can see that reasons-responsiveness is clearly impaired. Therefore, a person's habit of stealing is compulsive and unfree if the person is often unable to do otherwise in spite of good reasons to do otherwise that the person is aware of (at least aware of for some of the time, as the compulsion might cloud their judgement).

Nevertheless, the thesis that I suggest in this section is that freedom requires that the free person not be bound by *any* deterministic too-high-level laws – or generalisations that come too close to being deterministic laws. This is another reason it is unsurprising that the causal mode of thinking is incompatible with the intentional mode of thinking.

5.4.3 Flexibility and freedom

So why does it make sense to think that no deterministic law that is on a too high level can apply to free actions? And what does "too high level" mean?

The world is constantly facing us with new situations with new properties. The features of the world endlessly join together to create new combinations that often cannot be dealt with the same way as previous ones. Because of this, successful goal-directed activity cannot, in principle and often in practice, be bound by any causal generalisations. Further, it cannot be bound by any high-level, non-vacuous teleological rules either.

Imagine a very simple situation in which you are dropped into an empty field and must move so that you reach a physical location at its centre, say marked with a flag. It is easy to see what way you need to move to get to the flag as fast as possible, no matter where you are dropped – but what you must do is defined by your goal, not any rule made up beforehand. If you are dropped one hundred metres south of the flag, you need to move a hundred metres to the south. If you are dropped a metre

to the south-west, you need to move a metre to the north-east. If you are obliged to always move fifty metres to the north, there is only one spot from which you can end up by the flag by doing that.

But then, "move in a straight line towards the flag until you reach it" is itself a kind of rule – containing some flexibility and sensitivity to the conditions, but only some. It only works on the open field. Suppose you are in a labyrinth instead. Even if you know exactly where the flag is in the labyrinth and what the labyrinth's layout is like, you likely cannot get to the flag by walking in a straight line. You might even have to backtrack a lot first. So, following the one rule will not get you to your goal in this different kind of situation.

Then again, suppose you can easily get to the flag, but if you take the time to do that, you miss the chance to save a person having a heart attack by the field – and you care more about saving them. Then, even "Reach the flag by whatever means works best" is too much of a rule. And you cannot amend this by making up a rule beforehand that says, "Reach the flag by whatever means works best unless someone is having a heart attack by the field," because what if it is something else than that happens? If you make the rule say "Reach the flag by whatever means works best unless there is something more important to do," that would in some sense be the same as saying "Go to the flag or do not go," and it would not give guidance as to what to actually do.[319]

Because we are always confronted with new situations, we cannot react to them by any set of preset rules guiding action. We need to creatively come up with answers to novel questions about what we should do in a particular situation. This involves both seeing what will get us to a particular goal and what goals are important in what situation. This does not contradict determinism or causality. We can, and probably do, do things like model the situation in our minds, see what the

[319]To a human with an existing decision-making faculty the "unless there is something more important" version could give usable advice, but the person would then be following other rules of decision-making, not just the rule given.

consequences of different actions would be, and what kind of actions would lead to the desired consequences, and this is not indeterministic (or at any rate, we stand to gain little if it is – cf. 3.4.9). However, it is something that we cannot do if our actions are guided by rules that make any kind of non-teleological generalisations about what we will do in certain circumstances. (Recall the stereotypical addiction again.) If those circumstances are described as generalised, that is, through higher-level descriptions, that means they will apply in many possible individual circumstances. Those individual circumstances will also be different from each other, as all individual circumstances are, and less trivially, they may be *relevantly* different from each other. Thus, a rule prescribing a certain choice in certain kinds of circumstances may end up prescribing an action that is not conducive to the realisation of one's goals. Following a higher-level rule potentially means being unable to reach things that are valuable to us.

Nevertheless, it is important to notice the "may" and "potential". Trying to choose the most efficient or rational or optimised strategy in each situation is not always practically feasible. Going by a number of simplified rules – heuristics, rules of thumb – can often be more conducive to realising one's goals. They are a two-edged sword; they can both help and hurt us.

An important question remains unanswered: What level is low (detailed) enough, yet at the same time high enough? If we cannot have determinism on a higher level that is "too high" and be free, but also need it on some level in order to be free, when will we get to a deep level of description where determinism no longer threatens freedom? The argument in the previous paragraphs seems to lead to a regress that makes every level except the ultimate level too high. (The regress stops by the ultimate level because, in an ultimate-level description, there is nothing left that is not taken into account and thus, no overly broad generalisations about kinds of situations can be made.) That is not really the case, though. Something like the description of the decision machine above will suffice for freedom. If we want a "lawful" description that avoids the regress, we can state that the agent always takes into account all of their interests appropriately in making a decision. This is not a

deterministic description because it is too vague. To arrive at a deterministic set of laws for the agent, we would need to describe a sort of complete program for a decision machine – something that would be so free of ambiguity that it could be a computer program allowing a computer to actually be or run a decision machine. This would avoid the regress because it would make lawlike statements about how to make decisions based on any unique input, not how to behave in reaction to any finite, pre-stated conditions. I will call this level (that is, any level of description matching these criteria) the *deep level* for simplicity.

Note that the above brings out an assumption that it is possible to create instructions for a perfectly reasons-responsive decision machine. It is beyond my ability to determine whether this is possible in principle. If it is not, then by these standards, compatibilism is not quite true because it is not possible for a decision machine to be *perfectly* free.[320] This does not worry me much, for two reasons. Firstly, it is still possible in principle to be an extremely good reasons-responsive decision machine. Secondly, we are not in actual fact even good decision machines (see 9.6), so the theoretical possibility of absolute perfection is not a very relevant question.

In chapter 3, I argued that we should adopt the view that freedom requires determinism because this is required for things about freedom that are independently important for us. Now, I argue that we should adopt the view that freedom implies reasons-responsiveness for the same kind of reason: because without it, we may also end up doing things we do and should not wish to do (and because some existing examples of freedom and unfreedom agree with this). A corollary to reasons-responsiveness is that it excludes the possibility of following any higher-level (higher than deep-level) laws. Thus, combining these conclusions, we can say that freedom requires determinism on such a level that it enables us to reliably make decisions that take our interests into account but requires indeterminism in terms of

[320]Cf. Dennett 2015, pp. 32–35.

any higher-level description. We must be like (some variant of) the decision machine to be free.

In chapters 8 and 9, it will also be seen why it makes sense to make this a requirement for being held responsible, and why there is a (meaningful, non-primitive) connection between freedom and responsibility.

5.4.4 The choice function and the set of options: A working version of Christian List's theory

This section examines another perspective on partial indeterminism on a higher level based on developing Christian List's theory further. It summarises an argument I make in an article I have so far published as a preprint.[321]

Recall that List's "compatibilist libertarianism" is based on determinism on the ultimate level being compatible with free will if free choices are indeterministic on the higher psychological or intentional level.[322] As I argued previously (3.4.3), this gives no help against the randomness argument: the first main claim applies in that this indeterminism on the higher level compromises concrete control, and the second main claim stands in that List (at least sometimes puts the matter as if he) merely assumes that indeterminism is needed, but it really brings nothing desirable. In regards to this latter point, as I will show below, List also makes some arguments similar to van Inwagen's point about belief in determinism making deliberation impossible (see 3.4.9), and those can be answered in the same way, still not giving an independent reason to desire indeterminism. However, if we look at what List says more closely, and if we then apply the ideas from earlier in the present chapter, we can come up with something that List does not exactly say but that builds upon some of what he does say.

List does have an answer to the randomness argument at this point that is of a

[321]Kokko 2023.

[322]List 2019b.

different sort than what has been discussed so far (that is, mainly in 3.4), but it does not bring in anything very new as such. However, it can be used as a stepping stone towards a more detailed (and more compatibilist) theory. List develops his answer to the randomness argument with Wlodek Rabinowicz. They conceptualise it as answering two different intuitions at the same time, a version of the dilemma between control and open options that has been coming up in different forms above. One of the intuitions requires alternative possibilities, and the other one requires endorsement in the sense that the choice made must be picked in the right way, not in just any undetermined way.[323]

To answer these requirements, List and Rabinowicz posit a difference between options that are open to the agent and options that the agent can pick with endorsement. What options can be endorsed depends on the agent's motives; something that is against them cannot be chosen with endorsement. Since endorsement is a requirement for freedom, if the agent were to choose an option that is counter to their reasons, that choice could not be free. Thus, to briefly explain the model List and Rabinowicz propose, there is a set of possible options on the one hand and an *endorsement function* on the other, with the latter determining which of the options can be endorsed. They also use this to explain the example of Martin Luther saying he could do no other when he was doing what he thought was right – the one that has been used as an example of how someone could be very much responsible while not being able to do otherwise (mentioned in the present work in 2.4.2). They interpret him to mean he could do nothing else while still maintaining his integrity. In other words, it is not that he had no other options open, but this was the only one he could endorse.[324]

While this theory has much potential, at least with modifications, it has serious problems, too. Addressing which options can or may be chosen with endorsement

[323]List & Rabinowicz 2014, pp. 155–157.

[324]List & Rabinowicz 2014, pp. 155–160.

does not answer the questions of which actually can be chosen. Once again, we need to ask whether indeterminism or determinism apply to the choice, as well as where in it if so. First, List and Rabinowicz never say that it is not possible for someone to choose options not endorsed by the agent. If it is possible, then indeterminism certainly holds – as List states that it should on the higher level – but the randomness argument is not avoided. It is still the case that indeterminism may mean a loss of concrete control, and all the idea of endorsement adds is that if this risk happens to be actualised in the form of a non-endorsed option being chosen, then that choice turned out in a way that meant that it was not free, though we could not tell beforehand that it was not going to be. If, on the other hand, it is not possible in a normal choice situation for the person to choose against their endorsement, then the choice is really either deterministic or only indeterministic between good options.[325] There is nothing here that gets us out of the dilemma that either the choice as a whole is determined or it is (concretely) uncontrolled.

However, there is a way forward here provided we are not obliged to stay incompatibilist. List has stated that humanities and social sciences assume indeterminism in human behaviour in that they assume an agent can choose between different options.[326] However, this is not really correct. Theories about decision-making assume the agent has options open somehow, sure, but then they try to predict what the choice will be. They do not just take it to be random. This is even seen in a specific model by List and Frank Dietrich[327]. Similarly, if as I pointed out above the choice function either does not determine behaviour or does so and is potentially

[325]If some endorsed options are still better than others, then if the choice is undetermined between endorsed options, we can still choose a relatively worse option, implying loss of concrete control. If (which is unlikely) all endorsed options are always equally good ones, then we have indeterminism of the sort that does no harm because it makes no difference to how well our interests are realised.

[326]List 2013, p. 168.

[327]Dietrich & List 2016.

deterministic – why not go for the option that it *is* deterministic? This makes for a model that contains both the element of control (or endorsement) and the open options as a separate component, but it avoids the randomness argument by being overall deterministic. The level of description only involving the endorsement function – or choice function, since it does more than merely endorses – is deterministic, and the function itself is deterministic. The level of description only containing the set of options is indeterministic, not constraining the choice function from the outside. The combination of the two is deterministic, guided by the choice function.

The basic idea is the same as when we countered van Inwagen's notion of deliberation as involving belief in options being open: yes, they need to be open to deliberate between, but that is not properly analysable as indeterminism. Here, we see how the same thing appears in the light of how choices are analysed in fields studying them. There is a part of the system that needs to have the choices open, but the whole thing overall should not be indeterministic.

5.4.5 Inflexibility and the mode of causality

We can now return to the three modes of thought from before and see another side of why the mode of causality contradicts the intentional mode and our view of freedom, even aside from the reason that this was postulated in their properties. If we think about an agent as reasons-responsive in one of the three modes, it has to be the intentional mode. The mode of randomness is excluded for the same reasons as why the indeterminism that it effectively models makes no sense as a requirement for freedom, as per chapter 3. More significantly, the causal mode will not work, because the causal mode consists of mechanical, lawful explanations that are essentially never on the deep level. (Of course, for reasons previously discussed, even applying the causal mode on the deep level would probably make agency *seem* to vanish.) Explanations in the causal mode are of the form that something causes something else because of a law-like generalisation; because this generalisation is basically never on the deep level, causal mode explanations do not allow for reasons-

responsiveness. I think that this is a major reason why confusing determinism with the causal mode could make it seem as though determinism contradicts freedom. Further, even if we ignore the idea of the three modes, we can say that the observation that higher-level deterministic generalisations contradict freedom can explain the appearance that determinism does. This is also why freedom tends to contradict predictability, as seen in the prediction machine thought experiment.

5.5 Conclusions to chapter 5

This chapter has served to intertwined purposes: to make intuitions behind incompatibilism more explicable after compatibilist conclusions have already been established, and to begin looking at what freedom requires besides determinism.

5.5.1 Understanding “incompatibilist” intuitions

I cannot prove that the intuitions of people in general follow the logic I have proposed in this chapter. The statement is probably too generalising to be true anyway. However, I can say this much: there are a lot of details that we can find about different people’s intuitions that fit my proposal. If free will needed to be fundamentally deterministic, but in the right way and not in a number of wrong ones, would we not see what we are seeing now? People seeming to want indeterminism and determinism at the same time, indeed somehow wanting indeterminism but also wanting the concrete consequences of determinism, whether we are talking about contradictory layman responses or philosophers theorising about how to fulfil both the conditions of open alternatives and of control. If compatibilism in general seems to fit at least half our intuitions, my elaboration of it as reasons-responsiveness is as good as explanation as any for the totality of the mess of intuitions we do have.

Nevertheless, I can only take a vague stab at explaining where intuitions seemingly implying incompatibilism empirically come from. However, my observations here have opened a perspective that makes them make sense, and that is unlikely to be *completely* unrelated to where they come from. This is the idea that

can be simply summarised by saying that there is indeterminism – which is unfree – and there is both free and unfree determinism, but free determinism may be confused with indeterminism because unfree determinism can easily be confused with determinism in general.

From a slightly different perspective, the idea I have presented here is that the intuitive desire for indeterminism is really a perversion of a desire for reasons-responsive, open-ended, nuanced determinism where one's choices or decisions are determined by the whole picture in the specific situation – and the intuition of wanting to avoid determinism is a distorted version of a well-grounded desire to avoid higher-level determinism where choice or decisions are rigidly determined by only part of the situation.

I am not saying this is what the intuitions are "really" about. That statement would be both vague and overconfident, perhaps elitist, and unless I circumscribed its meaning in a suitably narrow way, I would have no tools to begin to prove such a claim. What I can say, and do in fact argue to be the case, is that interpreting those vague intuitions in this way would make a whole lot of sense, solve a lot of problems, and open fruitful avenues for further investigation. I also claim that the intuitions almost certainly have *something* to do with what I am saying, since my suggested explanation is intimately tied to the issues of freedom, alternatives, determinism and so on that the incompatibilist intuitions also have to do with.

So, never mind where exactly the intuitions come from (which would be hard to know) or what exactly their contents are (which would on top of that be hard to even define). What I am really suggesting here is that I have been able to use examining them as an inspiration for finding a way forward, for a better new theory explaining the same things. We can certainly speculate that this has something to do with the origin of the intuitions, but we cannot know and need not assume that.

Recall the quote from Kane referenced in at the beginning of the chapter 3: "Many kinds of freedom worth wanting are indeed compatible with determinism. What we incompatibilists should be insisting upon instead is that there is *at least one*

kind of freedom worth wanting that is incompatible with determinism."[328] Mirroring this, though in a slightly less conciliatory tone, I can say this: I think those of us who believe in free will that is compatible with determinism should concede this much to our incompatibilist opponents: many kinds of determinism are indeed incompatible with the kind of freedom that is worth wanting. What we compatibilists should be insisting upon is that there is at least one kind of determinism that is compatible with the kind of freedom that is worth wanting – and indeed necessary for it.

5.5.2 The ability to do otherwise equals reasons-responsiveness equals concrete control

The randomness argument raised the problem with indeterminism compromising concrete control. This chapter has moved into the direction of replacing the idea of indeterminism in general as a requirement for freedom with higher-level indeterminism combined with deep (psychological or intentional) level determinism, and I have argued that this points to the importance of reasons-responsiveness for freedom: being able to do otherwise when there are good enough reasons (that is, simply *doing* otherwise, without needing to wonder about what "being able" means), and being able to *not* do otherwise when there are good enough reasons *not* to do otherwise.

This takes us a full circle. Recall that the kind of concrete control that would be compromised by deep-level indeterminism meant being able to rely on following one's own reasons. This quite plainly refers to some kind of reasons-responsiveness.

We have seen that control and alternative possibilities are two important sides of free will that seem to contradict each other on a closer inspection. Compared to this, reasons-responsiveness brings together a less than obvious but rationally desirable version of alternative possibilities with a not very surprising and again quite desirable concept of control.

[328]Kane 2002a, p. 223, italics in original.

5.5.3 Moving on to responsibility

Throughout the previous chapters, the question of responsibility has repeatedly popped up and been suppressed for the time being. Notably, I have declined to conclude that, since we can explain freedom or agency in a compatibilist sense, we therefore get compatibilist responsibility for free, too. This may follow from typical assumptions, but my way of argumentation in this work is never based on appealing to typical assumptions alone.

Nevertheless, it has been noted that responsibility is seen as being related to the freedom and agency, and that people tend to feel as though it is similarly threatened by incompatibilist arguments. In the following chapter, we will start to look at this challenge in more detail.

6 Moving from Freedom to Responsibility – and Punishment

The topic of this dissertation is defined as being about both free will and responsibility as well as their relationship. At this point, we have discussed free will in chapters 2, 3, 4, and 5. Chapter 2 introduced the problem and chapters 3 and 4 presented the randomness argument against libertarian free will and the argument that there is nothing meaningful lost in a compatibilist conception of freedom. Chapter 5 went on to examine the special nature of freedom as an application of determinism, drawing on insights gained from the apparent incompatibility between determinism and freedom. It is now time to turn to responsibility. This chapter introduces concepts related to responsibility and starts, again, with incompatibilist intuitions about what responsibility is and requires.

There are different possible ways to approach the connection between free will and responsibility. I have already been formulating my theory of free will based on what concrete consequences different views have. For the ethical questions concerning free will and responsibility, I want to focus on the concrete consequences of these ideas in the sense of what kind of action they guide us to take and what consequences these have in the world. For some of the arguments in this work, an important aspect of the question of justifying the connection – and justifying the very notion of responsibility as we understand it – is the question of the connection between responsibility and punishment. For this reason, I start approaching the concept of responsibility by examining ideas that tie it to punishment.

I look at responsibility through two thought-provoking and in many ways excellent books that I nevertheless disagree with about some fundamental things.

Göran Duus-Otterström's dissertation *Punishment and Personal Responsibility*[329] examines many aspects of the concepts mentioned in its title and makes an argument for a retributivist theory of punishment, and I will summarise quite a few parts of it because they are all relevant to understanding just what responsibility is. The choice of this book as a focus is also based on its explicitly tying responsibility to libertarian free will. Meanwhile, Saul Smilansky's *Free Will and Illusion*[330] shares some of Duus-Otterström's basic assumptions about the meaning of freedom and responsibility, but it combines them with an awareness of the randomness argument and ends up going in much wilder directions. I will summarise the basic argument of this book. After going through these two books, we have an idea of the questions and assumptions surrounding the concept of responsibility.[331] I then go on to ask more questions and question the assumptions, setting the stage for the next chapter, where I dissect the concept of responsibility.

6.1 Göran Duus-Otterström: Punishment and Personal Responsibility

Duus-Otterström's main questions concern the justification of punishment. He mentions that punishment seems morally problematic considering that it involves doing intentional harm to others, something normally seen as morally unacceptable[332]. Nevertheless, he immediately goes on to state that he is going to

[329]Duus-Otterström 2007. Below, in section 6.1, which discusses Duus-Otterström specifically, this book will be referenced with bare page numbers. No other works are referenced in 6.1.

[330]Smilansky 2000. Below, in section 6.2, which discusses Smilansky specifically, this book will be referenced with bare page numbers. No other works are referenced in 6.2.

[331]For other overviews of theories of punishment see e.g. Bean 1981, Hart 2008, Yli-Hemminki et al. 2022, Honderich, 2002 pp. 135–141. I am making a conscious choice to keep to the classifications and theories given by Duus-Otterström, which are sufficient for the purpose of this study.

[332]Duus-Otterström 2007, pp. 8–9.

assume that punishment is justified, some way or other[333]. The major question he will be asking is stated as

> (Q) Which principle or theory (or principles or theories) should serve as the basis for a state's penal regime?[334]

In my estimation, this is on one hand a good way of defining a sufficiently limited question to be answering in a single work, but on the other hand, a way of asking the question that takes too much for granted. I will return to my criticism below.

Duus-Otterström's answer to the question is that what he calls *retributivism* is the answer to *Q*, at least compared to those rivals he considers. This is one of the many concepts in need of a definition that Duus-Otterström provides handy definitions for.

6.1.1 Useful definitions

Duus-Otterström gives definitions to many terms related to punishments, responsibility and desert that are useful and necessary not only for understanding his work but also in talking about the topic of responsibility in general.

The central concept of *punishment* is defined by Duus-Otterström[335] as:

1. Pain or deprivation (or what is intended to be felt as such, even if it is not)
2. intentionally inflicted
3. on a person (here seen as meaning a human individual)

---

[333] 9.

[334] 9.

[335] 48–50.

4. by an authority
5. for an offence (which is explained to mean that the punisher has a particular offence in mind and at least thinks the punishee is guilty for it, ruling out the possibility of using the term *punishment* for anything intentionally done to the innocent).

One concept opposed to that of punishment is *treatment*, which here means steps taken against a rule-breaker by an authority that are not meant to cause pain or deprivation but to treat the disorder seen as leading to their breaking the rules[336]. Naturally, *treatment* in basically the same sense can be used for steps taken to help someone in general to recover from some kind of disorder, something that will come up later in the present work (8.2.3).

Duus-Otterström also clarifies what he means by wrongdoing in the relevant sense, i.e. in practice by *offence* above. Not all rules justify punishing those breaking them. Particular rules could even be morally incorrect themselves, for example. Duus-Otterström's discussion concerns offences that it makes sense to criminalise, are in fact criminalised (this is not stated but seems implied), in other words there are rules against them, and breaking those rules is *prima facie* morally wrong. Further, "everyone" can agree about this *prima facie* wrongness. This includes even perpetrators of offences against these rules – they will agree that breaking the rules is *prima facie* wrong, that is, even if they do not agree about their particular offences being wrong. Only such rules, breaking them, and punishment for breaking them are to be considered, and it is assumed that, in a society, there are at least some such rules.[337] Since the requirement that *everyone* agree on something is very easily falsified, I think it is better to think of the arguments as applying when a large majority of normal adult people would agree.

[336] 50.

[337] 56–57.

The question that Duus-Otterström asks concerns the bases of a *penal regime*, not just punishment. He explains his concept of a penal regime as follows:

> A penal regime, as I understand it here, is the whole set of formal practices and rules the state employs with respect to lawbreaking. Punishment, and the reason for inflicting it, takes central stage in such a regime. But a penal regime also specifies when *not* to punish. It includes criteria for when to hold responsible and when to excuse, as well as considerations on the forms and purposes of penal responses.[338]

Excusing, mentioned above, is also a significant concept. Duus-Otterström discusses three reasons for "withdrawing a punishment that seems *prima facie* appropriate"[339]: *excuses*, *justification*, and *mercy*. An excuse is a reason to think that the offence was not the person's fault – accepting the wrongness of the act, denying the agent's responsibility for it.[340] A justification is a reason to think the alleged offence was not in fact blameworthy – accepting responsibility, denying wrongness. Mercy is given when neither of the above holds (accepting responsibility and wrongness), so punishment is deserved, but it is withheld anyway.[341] It is worth noting that "could not do otherwise" is a paradigmatic form of the excuse, just as it is of lack of freedom.

---

[338] 58.

[339] 207.

[340] *Excuse* is not to be confused with the Finnish *tekosyy*, which means a contrived excuse. Excuses can be perfectly good ones, even if the word "excuse" is often used in such a context as to be there translatable as "tekosyy", which is not the case in the present context.

[341] 207–209. Duus-Otterström mentions that the possible fourth case of denying both wrongness and responsibility is unlikely to come up. I could imagine using it if e.g. being accused of being homosexual: in cases where people fundamentally disagree on what kind of things are morally wrong in the first place. Duus-Otterström briefly discusses this in 224, almost making the same point.

Finally, we have the three theories of punishment that Duus-Otterström compares in his book. According to *deterrentism*, the purpose of punishment is to deter future criminal acts. According to *rehabilitationism*, criminals are as it were sick and need to be treated, and the purpose of "punishment" is just this. As we can see from the definitions above, this really counts as treatment and not punishment. Lastly, according to *retributivism*, punishments should be handed out because the perpetrators of offences deserve them based on what they did.[342]

6.1.2 The argument for retributivism over deterrentism and rehabilitationism

Duus-Otterström mainly compares retributivism to the two other theories of punishment defined above, noting that there are other possibilities but these are the main plausible rivals.[343] He argues that retributivism is superior because it is more just (the institutional reason) and because it treats people respectfully as free and responsible agents (the symbolic reason).

As for retributivism being more just, Duus-Otterström tries to skirt around being circular here but is not very successful. After all, his retributivism could almost be alternatively defined as handing out punishments if and only if it is just. In any case, with justice as a standard, retributivism stands out as a deontological theory against the utilitarian bases of the two others, and the objections against the utilitarian theories are some of the classic ones. It might be useful for the sake of deterrence to "punish" the innocent, or not punish the guilty. Treatment-style "punishment" might be usefully applied to those who have not done anything. The "punishments" in either case might be more effective for deterrence or rehabilitation if they were not of a length or severity proportionate to the offence. All of this would be unjust, and none of it would be allowed under retributivism.[344]

[342]Chapter 3.

[343]61.

[344]Chapters 5 and 6.

Utilitarians can claim that unjust procedures such as "punishing" the innocent would not in fact maximise utility, but this, to Duus-Otterström, is not the point. The point is that utilitarianism may allow such things in principle and thus is not just.[345]

However, there is an oddity with this objection in conjunction with something else Duus-Otterström says. I will not go so far as to say the objection is thereby invalid, but this oddity is worth noticing here, and I will have more to say in a similar vein later. The oddity is that desert-based retributivism as explained by Duus-Otterström does not always sound so different in this respect. Duus-Otterström is willing to compromise on everyone getting their just punishments if there is a strong enough (utilitarian) reason[346]. Thus, while Duus-Otterström condemns utilitarian "justice" for allowing injustice in principle even if it never allows it in practice, he at the same time accepts his own deontological model allowing injustice in practice as long as it condemns it in principle. To put it a little nastily, he seems more concerned about what we say than what we do. One way to look at why this happens is through a perspective that sees utilitarian and deontological principles as less opposed to each other than their traditional formulations suggest; I will have more to say about such an idea in my own positive theory of responsibility in chapter 8. In any case, I agree that ethics in real life needs both sticking to principles and not applying them too strictly, but this sits ill with being strict about rejecting a rival ethical view based on its lack of principled strictness.

The symbolic reason for favouring retributivism is perhaps more interesting. According to Duus-Otterström, punishing people based on their desert affirms that they are treated as responsible persons.[347]

In outlining the symbolic reason for retribution, Duus-Otterström presents

---

[345] 140.

[346] 146.

[347] Chapter 6; the rest of this subsection is all based on that chapter.

three models of rule-breaking associated with the three models of punishment: the *disorder model*, the *autonomy model*, and the *rationality model*.

The disorder model is associated with the rehabilitation model: rule-breakers are seen as disordered: sick, abnormal, not in control of themselves. That is why what they need is treatment. This does the opposite of affirming that they are free and responsible agents.

The autonomy model is associated with retribution: rule-breakers are seen as choosing freely and autonomously and being responsible for their choices. This is why they are punished as ones responsible for what they have done. Justice-based punishment implies that its objects are free and responsible.

Unlike rehabilitation and retribution, the deterrence model is not associated with one model of rule-breaking. Both disorder and autonomy are compatible with it. However, according to Duus-Otterström, the third model of rule-breaking, the rationality model, is exclusive to deterrence. In this model, persons act so as to maximise their own utility, so it makes sense to weigh the scales with the threat of punishment – the very idea of deterrence.

The problem with the rationality model is its determinism, whereby agents are assumed to always choose the option that has the highest expected utility, or one of those options if there are several equal ones. Indeed, even though my own theory will be based on determinism and rationality, I would tend to agree such a simplistic model of rationality with such simplistic determinism contradicts freedom (see 5.4 in the present work).

Thus, Duus-Otterström has argued that retributivism is preferable to the other two theories for two reasons: it is more just, and unlike them, it treats people as responsible agents.

6.1.3 Two objections to retributivism

Duus-Otterström discusses two objections to his kind of idea of retributivism, both

of which bear mentioning here.

In his chapter 7, Duus-Otterström discusses the idea that people are not free and responsible because science can explain their actions in various ways. This is familiar territory from chapter 5 of the present work (though the question of what science actually says will be discussed in section 9.6.1). Duus-Otterström argues that ultimately any kind of explanations of human action are excusing, that is, give grounds to think that the person is not responsible because their actions are ultimately explainable by factors beyond their own control. Only being a *causa sui*, in other words libertarian free will, can make a person responsible, making their acts not completely explainable in this sense. It is now easy to respond to this by saying that it is really a question of looking at the person's actions and choices under the different modes of causation and freedom introduced in section 5.3 of the present work. Just because explanations exist does not mean that we cannot also apply the perspective of freedom. What has not been discussed is how this relates to responsibility and whether responsibility would require a different kind of freedom. That will be discussed in the following chapters of the present work, where I will also discuss why there should be a connection between freedom and responsibility in the first place.

The second objection, discussed in Duus-Otterström's chapter 8, is simply the thesis of hard determinism. The argument is familiar from our earlier discussion of incompatibilism (2.4; see also 4.3.5). If people are determined by factors outside their control, as will always happen in some sense when universal determinism or just determinism about human actions holds, then they cannot do otherwise, are not free, and are not responsible. Duus-Otterström's response to this is to appeal to the fact that we do not know whether it is true, and we should bet that it is not.

6.1.4 Conclusions: Betting against hard determinism (and betting for what?)

Duus-Otterström argues that we do not and perhaps cannot know whether determinism is true. He concludes, however, that since we do not know, it is best for

us to act as if "libertarian free will" exists, and to hold people responsible, maintaining the state's penal regime. Put shortly, his argument is that a world in which freedom and responsibility exist (and we act accordingly) is preferable to the other alternative, and there is less to lose by betting that we live in such a world than by betting on hard determinism. It is not that it would be better to pretend freedom and desert exists even if (we knew) they did not; the bet is meant to have better expected outcomes from our current and perhaps permanent state of uncertainty.[348]

Duus-Otterström also outlines the basic properties that a penal regime based on desert should have. In this model, penal responses are meant to inflict pain or deprivation on a person based on their deserving it, not to rehabilitate or deter, though these can be regarded as welcome side-effects. They should reflect some kind of proportionality of crime and punishment, though the details are much up for debate. They should never be pre-emptive, "punishing" someone who has not done anything yet. They should be sensitive to genuine justifications and excuses. More interestingly, all punishments should be of a determinate length and determined by the seriousness of to the offence. This means that the severity of punishments (the length of prison sentences being the prime example) should not be affected in any way by other considerations such as the opinion of the victims or whether the criminal is a first time offender. Is also means that the punishments should not be altered afterwards for good or bad behaviour.[349]

When Duus-Otterström goes through all the caveats to his "calculation" about which is the safer bet (Are the odds for determinism really 50–50? Is betting on freedom when there is none a very bad outcome because of the injustice involved in punishing the non-responsible? etc.), he finds enough[350], in my judgement if not his

[348]Chapter 9.

[349]333–337.

[350]315–318.

own, to undermine this whole argument. The whole thing seems to collapse into uncertainty; we do not even know which bet is really safer with any reasonable degree of confidence. However, for the sake of my own argument, I will ignore these problems. I will take it that, with his own assumptions, Duus-Otterström can make a case that we should bet on freedom (or "libertarian free will") rather than hard determinism because the existence of freedom is such a desirable state. I will contest this idea on a different basis.

One of my two arguments against Duus-Otterström's bet is that he has not actually established that the connection between desert and punishment makes sense and, for all that the connection is often taken for granted, there is a very serious ethical question about it. I will return to this topic later.

My other objection is that Duus-Otterström accepts the notion of libertarian free will as true freedom – the idea I rejected via lengthy argument in chapters 3 and 4 of the present work. There is nothing about his conception of what freedom is supposed to do that would cause those arguments to fail to apply. To him, freedom is something valuable that implies autonomy, control, and responsibility. Autonomy and control cannot be gained from randomness, and intuitively, it seems responsibility cannot be either.

Duus-Otterström does discuss theories of free will extensively[351]. If one is not looking for it, his way of ignoring the randomness argument might be inconspicuous. When looking for his take on the argument, his ignoring it becomes quite conspicuous instead. He does *mention* the randomness argument, but the closest he ever comes to *answering* it is making brief supportive but inconclusive remarks for a third option that is neither "determinism" nor "randomness"[352]. Thus, he begins to

[351]Especially chapter 8.

[352]270. For discussion related to this topic in general in the present work, see 2.2.2, 3.2.1 and 3.4.5.

acknowledge the problem, but he does not take it seriously enough for it to affect his main line of argumentation. This is an example of why the randomness argument should be taken more seriously: it is, as I have argued, objectively strong enough to undermine libertarianism and even incompatibilism seriously, but when it is treated merely as one of the threads of the existing discussion to be mentioned for the sake of completeness, one can end up building an otherwise beautiful logical construction standing on foundations that are really nothing but air.

So, what happens if we take basic assumptions about the nature and connection between freedom and responsibility that are similar to Duus-Otterström's but, at the same time, acknowledge the power of the randomness argument? Such a view is presented by Saul Smilansky, to whom we turn now.

6.2 Saul Smilansky: Free Will and Illusion

Saul Smilansky's book *Free Will and Illusion*[353] argues that it is necessary for most people most of the time to have an illusioned view of what free will is and can be like. The reason he gives for this is the following: Only what he calls "libertarian free will" would be complete freedom and allow responsibility in the fullest sense, yet such "libertarian free will" is impossible. However, freedom as understood in a compatibilist sense is also real – a partial freedom, leading to partial responsibility, important yet incomplete. It is important to respect freedom and responsibility in the limited compatibilist sense in which they do exist. However, there are strong reasons why awareness of the impossibility of ultimate freedom and responsibility might prevent a person from respecting the limited freedom and responsibility that are possible. Therefore, to ensure that what freedom and responsibility there are will be respected, it is usually desirable for people to have the illusion that freedom and responsibility exist in a stronger sense: either that libertarian freedom is possible, or that compatibilist freedom is freedom in the full sense.

[353]Smilansky 2000.

Smilansky's approach is bold in a good way, and his way of acknowledging the different perspectives on free will parallels my own approach in my dissertation. Nevertheless, I find that his conclusions are too bold because his starting point is not bold enough. Illusion is not important; different perspectives are, and understanding them removes the apparent need for illusion.

Smilansky's argument involves many terms of his own devising, so it will be convenient to introduce many of the themes he discusses by starting with the terms. This introduction to his argument is mainly based on the first part of the book, not so much the second, which explores the consequences of the conclusions of the first part.

A central notion is what Smilansky calls the *Core Conception*. This is a basic moral notion and intuition that states that stresses the moral importance of considering a person's control (also called *up to usness*). In terms of responsibility, this means that a person can be held responsible for something only insofar as it was under that person's control, or up to them, to make that thing happen or not. Holding someone responsible for something that they had no control over is, after all, a paradigm of injustice. The Core Conception is also essential for respect towards persons since acting is an important part of what being a person is all about. It is initially formulated neutrally in terms of compatibilism or incompatibilism.[354]

Smilansky rejects a purely utilitarian basis for holding responsible as being inherently immoral, against the Core Conception. That whether the innocent could be punished depends on whether it would bring good results goes against our basic moral intuitions and the basic values that Smilansky accepts. This also entails rejecting what Smilansky terms e*ffect compatibilism* – a view according to which it is right to punish someone when it produces good results, regardless of whether that would be just by more traditional intuitions. Instead, the kind of compatibilism he considers later on as plausible is *control compatibilism*: it is right to hold responsible

[354] 14–22.

based on free will and control in the compatibilist sense.[355]

Smilansky also criticises and rejects what he calls *the assumption of monism*: That it must be the case that either compatibilism is true or that incompatibilism is true, in answer to the compatibilist question. While it is true that the theses in their strong form are incompatible, it is possible to hold a mixed view in which parts of each of the two views are true.[356] Instead, Smilansky thinks that there is a *Fundamental Dualism*: there is a sense in which control compatibilism is true, ways in which we should be regarded as free and responsible; but there is also a sense in which hard determinism is true, ways in which we truly are not free or responsible, since there is an *ultimate perspective* from which it can be seen that we are not ultimately in control.[357]

Smilansky thinks that what he calls "libertarian free will" is what would satisfy the requirements for freedom on every level, but it is also impossible. Why it is not possible will not take much explaining: it is a version of the randomness argument. What Smilansky sees as the heart of "libertarian free will" is a person being the ultimate origin of their choices, and I will refer to this as full ultimate origination instead from now on, it being the same thing referred to by that name in 3.6 previously in this work. Smilansky agrees about the argument presented there: this is impossible because in order for a person to be an origin of their actions, the person has to exist and have properties already; but then the person must have been created by something else, and that something else will be a more ultimate origin of the person's actions than the person is. Indeterminism does not help either because it is mere randomness.[358]

---

[355]27–33.

[356]36–38.

[357]Chapter 6.

[358]Chapter 4.

Compatibilism cannot be the whole truth either because there is always the ultimate perspective. There is some sense, some perspective, in which persons cannot be responsible for what they do, since only with impossible full ultimate origination could they be responsible for everything that leads up to their choices. If determinism is true, there was always something before that determined every choice of theirs, and this something was itself not up to them. (If determinism does not hold, then some of these determined things are replaced by useless randomness.) Thus, compatibilism by itself is morally shallow, in not taking into account the deeper level, and complacently unjust, in dismissing the ultimate-level injustice inherent in holding people responsible when they cannot be ultimately responsible.[359]

However, hard determinism is not the whole truth either. If we were to conclude that there is no such thing as freedom or responsibility at all, we would miss out on the important ways in which there is. There is genuine, if partial, freedom that we need to respect. There is a partial justification for holding responsible and punishing, and there is a practical necessity for doing so. Though we cannot have complete responsibility, there is need to maintain a *Community of Responsibility* that respects the Core Conception in the control compatibilist sense that is possible.[360]

We are left with the result that full ultimate origination or "libertarian free will" is simply incoherent (Smilansky does not shy away from admitting this[361]), and both compatibilism and hard determinism are insufficient by themselves. Smilansky's conclusion is thus that the view that we must adopt is a mix of control compatibilism and hard determinism – hence, the Fundamental Dualism.

---

[359]Chapter 3.

[360]Chapter 5.

[361]"On the Coherence Question, the answer is simply negative. Logically, the conditions libertarians must pose for a worthwhile libertarian model of free will cannot exist." (73.) As for using the specific word "incoherent", on pages 48–50, Smilansky speaks repeatedly of the objection that "if" libertarian free will is incoherent, then it is not worth wanting. This is before his conclusion in the next chapter (chapter 4), quoted just above, that indeed it cannot be made coherent.

Because no-one is ultimately responsible for what they are like, and therefore for anything that they choose to do, there is an ultimate-level injustice in holding anyone responsible for anything. While it is true (and significant) that one can be responsible on the control compatibilist level, it is also true that there is a more fundamental level on which they are not. Thus, while we are compelled to deliver control compatibilist level justice to maintain the Community of Responsibility, even as we do so, we are committing injustice as well. On the ultimate level, someone who was the ultimate origin of their choices would be responsible, but there cannot possibly be any such person. Thus, (ultimate-level) injustice is a structural part of (control compatibilist) justice, and of life in general. We are missing something vital, and life is absurd ("absurd" is Smilansky's own wording[362]), yet there is enough partial moral value there that we need to respect it and go on living life with what meaning and morality there is. No wonder then that Smilansky thinks it is almost necessary to live in illusion about this.[363]

The illusion can be that compatibilist freedom and responsibility are entirely satisfactory, or that libertarian freedom (full ultimate origination) exists. Smilansky thinks that such illusion is already so prevalent that there is no need to promote it – it is merely right not to dispel it. Without this illusion, people would be prone to lose their sense of agency and worth, and also to neglect to maintain the important compatibilist-level sense of freedom, responsibility and respect. It would be theoretically possible, but hard, to live a moral life while being disillusioned; Smilansky does not even seem to count himself among such hypothetical individuals.[364]

[362]To be precise, "Life Is Absurd" is the title of section 11.5 in his book.

[363]Chapter 7. See also chapter 8.

[364]Chapter 7.

6.3 My response

As with defining free will, I refuse to accept the thesis that libertarian free will is necessary for responsibility and that we are left in a bad state without it. I have to take some care in how to conduct my rejection, however. Duus-Otterström and particularly Smilansky put an opponent in a strange position that is easy in one sense and difficult in another by appealing to intuitions that they are not willing to compromise on. What is an opponent who does not happen to feel the same way to do? One can reject their premisses as unfounded, but then people who do agree with those premisses will stop listening. What I will try to do here is to reject their positions by first discussing appeals to intuition to show why they should not be accepted here and then, mostly in the following chapters, explaining why there is a better case to be made for an opposing position anyway.

6.3.1 On appealing to intuition

> You say you cannot imagine that p, and therefore declare that p is impossible? Mightn't that be hubris? One of my tactics has been to respond to traditional philosophical claims about what is unimaginable by urging: try harder.
>
> –Daniel Dennett[365]

Philosophers often appeal to intuitions as part of proving some position. On some level, this is absolutely unavoidable. If we questioned everything that is some kind of an intuition, we could not even apply logic or make use of our senses.[366] However,

[365]Dennett 2015, p. 186.

[366]I am using *intuition* to mean any belief or inclination to believe something that comes to us directly, without an explicit chain of reasoning or evidence behind it. The idea is the same as expressed here:

> More promisingly, fans of intuition could narrow down the category by specifying that intuitive thinking is not based on a *conscious process* of interference. … Drawing the line between intuitive and non-intuitive judgements in that way has a significant

I think taking intuitions as indicative of the truth is often a mistake.

My stance towards intuitions is this: That a person has an intuition is a fact about the world, and the same goes for many people or even most people having an intuition. It should be taken into account the same way as any fact about the world. The existence of a given intuition may prove or suggest something relevant, but it does not, without some further argument, prove or even suggest that the content of the intuition is right. Suppose you see writing on a wall proclaiming that "Kilroy was here." Which are you going to conclude – that someone has been vandalising the wall to replicate a meme, or that Kilroy was there? The same goes with intuitions. With ethical intuitions, we might find it hard in practice to change them or get rid of them. We might even have reason to wonder about an ethical theory that goes contrary to our intuitions. However, to say that something is an ethical fact just because we feel so (or some of us do if they apply their intuitions in a certain way) is unwarranted.

Note that in spite of my critical stance towards intuitions, I do not necessarily accept the argument against thought experiments that they are inappropriate because they use too contrived scenarios[367]. I am more concerned with being careful about what a particular kind of proof can do in particular circumstances[368]. Thus, even in this section, I later look at some hypothetical scenarios to make my point. In this

---

result: *all non-intuitive thinking relies on intuitive thinking*. For if non-intuitive thinking is traced back through the conscious processes of inference on which it was based, sooner or later one always comes to some thinking not itself based on a conscious process of inference, which therefore counts as intuitive thinking. (Williamson 2020, pp. 54–55.)

It is not a given for "intuition" to be defined like this in philosophy (see e.g. Pust 2024), but I find it to be an appropriately simple definition and way of carving up nature. In addition, it seems to be what is common to different uses of "intuition", such as those used in both philosophy and psychology (see *ibid.*).

[367]E.g. Gasparatou 2008.

[368]Cf. Williamson 2020, p. 123.

work, any appeals to thought experiments or intuitions that I make will be of three sorts:

1. To appeal to intuition to demonstrate that a thought is *plausible*, without claiming it to be shown to be more than that (as e.g. 5.2.1).
2. To appeal to intuition in an attempt to counter *other* intuitions that potentially stand in the way of accepting the real argument, which is not based on appealing to intuition.
3. To use inference, not intuition, to demonstrate that, given a certain agreeable premisses, certain things will follow. A clear example of this last is right below when I demonstrate that Thomas Nagel's specific intuitions could be wrong without his being able to tell.

A warning example of an appeal to intuition as indicative of fact is found in Thomas Nagel's *Mind and Cosmos*[369]. The book as a whole has a problem with trying to override existing science with intuition[370], but the most blatant example is in Nagel's treatment of values. Nagel asserts that evolution by natural selection cannot explain our knowledge of values. He admits that natural selection can explain the existence of our feelings (i.e. intuitions) that things are good or bad; he uses pain as an example, so it is not hard to see how this relates to survival. However, Nagel goes on to say that his intuition that pain is a bad thing – an intuition which, I should reiterate, he has admitted would be perfectly explicable by natural selection – is so strong that he takes it to be objectively true.[371] Based on this, pain is objectively a bad thing in a way that has nothing to do with survival or, if we look at Nagel's

[369]Nagel 2012.

[370]See my review at <https://thoughtsonx.wordpress.com/2016/02/07/review-mind-and-cosmos-by-thomas-nagel/>.

[371]Nagel 2012, pp. 109–111.

theory of what values are[372], not much of anything else in the world.[373] Since this alleged objective fact has almost no connection to the material world, our knowing it cannot be explained by natural selection, since natural selection would be blind to such objective values. Since it is seen as a permissible move to appeal to intuitions in philosophy, Nagel uses an appeal to an individual intuition that he grants could be explained by natural selection to show that that same intuition cannot be explained by natural selection. Even though Nagel concedes such intuitions are not infallible[374], assuming they can even override arguments that *explain* those intuitions themselves is giving them an extremely high priority as evidence.

We can even construct a sceptical scenario from Nagel's own assumptions: supposing evolution was guided by natural selection (ignoring the other alleged problems Nagel sees in this, as presumably this argument is meant to stand on its own to some extent), then in such a world, creatures would have intuitions about something being good or bad that would not match objective moral facts. (Objective

[372]See chapter 5 in *op.cit.* as a whole.

[373]For other problems that would apply to such a theory, regardless of whether it is classified as "nonnaturalism" or not, see also this argument, as condensed in the abstract of Hayward 2019: "Non-naturalist realists are committed to the belief, famously voiced by Parfit, that if there are no non-natural facts then nothing matters. But it is morally objectionable to conditionalize all our moral commitments on the question of whether there are non-natural facts. Non-natural facts are causally inefficacious, and so make no difference to the world of our experience. And to be a realist about such facts is to hold that they are mind-independent. It is compatible with our experiences that there are no non-natural facts, or that they are very different from what we think. As Nagel says, realism makes scepticism intelligible. So the non-naturalist must hold that you might be wrong that your partner (for example) matters, even if you are correct about every natural, causal fact about your history and relationship. But to hold that conditional attitude to your partner would be a moral betrayal. So believing non-naturalist realism involves doing something immoral." Now, this is of course circular in assuming a different set of values than those it criticises, but just the possibility that it is right should raise alarms: we are taking a risk of being immoral in taking a risk of being wrong about a thesis that can imply that things do not morally matter.

[374]Nagel 2012, p. 110.

moral facts might exist in this world or not; in any case, they would not be knowable by creatures evolved by natural selection, by Nagel's own argument.) Further, such creatures could make the same kind of appeal to intuition to "prove" that their intuitions referred to objectively existing values. They, however, would be wrong. Only creatures whose intuitions would correspond to objective moral facts would be right in making such an argument.[375] Now, if you are a creature in either kind of world making such an argument based on an appeal to intuition, there is no way for you to tell which kind of world you are in, and hence, whether your appeal to intuition can be trusted. Thus, from your epistemic perspective, it cannot be trusted.

All of this is a lengthy way of making and exemplifying the point that appeals to intuition that assume the intuition tells the truth about something need to assume the reliability of the intuition, and thus are effectively circular. If you would believe an intuition, you should have a reason to do so. Certainly you should not have a reason *not* to do so, as with Nagel above.[376]

This is admittedly a very straightforward case of a faulty appeal to intuition. Duus-Otterström's and Smilansky's appeals are not as bad, but they suffer from the same general problem. As I will show below, both take some moral claims for granted. Further, since the intuitions are moral, they may be felt to carry a particular weight. You should not compromise on moral principles, after all, just because

[375]It could be said that a being who *in fact* has good reasons for its belief may have better justification for that belief than a being who is in an internally indistinguishable state but is so for the wrong reasons. (Simon Prosser brings this up in Prosser 2013, p. 322, citing the hard-to-contest example of hallucinations and experientially identical true perceptions, and labelling this view epistemological externalism.) However, this is something we can only rule from an external perspective where we already know the fact of the matter – whether we are dealing with hallucinations or real sensations, or intuitions created non-veridically by evolution or intuitions that are somehow mysteriously a source of real knowledge. There is no justification to make judgements as if we are in the justified position when we cannot actually know this (as Prosser also points out in the same source and on the same page).

[376]Cf. Häggqvist 1996, p. 123, which argues that there is little reason to trust intuitions just on the basis that they are intuitions.

someone uses some cold, utilitarian reasoning to try to override what you dearly believe. Or should you? Or should we be asking whether you should trust your intuitions of unknown origin so much that you give them the privilege to justify causing harm on purpose? I return to this soon in chapter 7, but I will say a few words here already.

We can give various arguments for accepting ultimate values, though because of the is–ought gap, they can never be conclusive. Given what I have said about intuitions here, appeals to intuition are particularly flimsy arguments compared to everything else. When we make a choice about *how* to make choices about ultimate values, why should we be satisfied with a method based on simply how we feel regardless of anything else?

Intuitions also have a weakness in that since, by definition, they are not arrived at by explicit reasoning, we do not know *what* they are arrived at by. Thus, we could be affected by factors we would not wish to be affected by. We can combat this by resorting to explicit reasoning to examine our intuitions... but not if appealing to intuition to override anything else is a permissible and even favoured move in philosophical investigation. While Duus-Otterström and Smilansky may not do this as blatantly as Nagel in that one example, if they remain resolute about their position about the importance of libertarian free will or full ultimate origination, they are forced to make that move.

After this detour into the limits of appeals to intuition, I now return to discussing Duus-Otterström's and Smilansky's ideas. Both of them face the issue that their starting point involves assuming a connection between responsibility, punishment, and "libertarian free will", which leads them to take positions they cannot justify without giving intuition a very strong role.

6.3.2 Questioning Duus-Otterström

Duus-Otterström speaks of a "gap in retributivism" that he needs to bridge. He formulates this as concerning why it is right to punish people if they deserve it. I

think this seems a little confused; what does someone *deserving* to be punished even mean if not that it is right or obligatory to punish them?[377] Nevertheless, there is certainly some such gap – to me, perhaps between wrongdoing and deserving to be punished instead. Duus-Otterström tries to bridge it without resorting to *intrinsic-good retributivism*, which would mean just taking it that punishment is good in itself in some cases.[378] That is why he brings up his institutional and symbolic reasons for punishment. Nevertheless, the gap remains.

Duus-Otterström takes rules as a given, saying that there are many benefits claimed for having them, so (basically) surely at least some of those really hold[379]. Punishment, in turn, is something that is applied to those breaking these rules, and thus, there would be no punishment without rules. This, taken together with the fact that punishment is seen as something of a primitive moral requirement, has an odd consequence: Rules are a necessary requirement for (the justification of) punishment and somehow primary to punishment, but at the same time, rules exist to further other more fundamental values, yet the justification of punishment is not based on furthering any other values. Rules are grounded in other moral values, rules ground punishment, yet punishment is not grounded in other moral values. (Though Duus-Otterström does not address this point, his discussion of whether crime is socially constructed[380] is marginally related.)

Though Duus-Otterström verbally rejects intrinsic-value retributivism, he ends up assuming something much like it anyway. Retributivism is defended on the basis that it is just, and on the basis that it respects persons. Why is it just? Because people get what they deserve, and what the deserve for wrongdoing is punishment.

[377]See also 7.1, 7.2 and 7.3 in the present work.

[378]Duus-Otterström 2007, pp. 107–116.

[379]*Op. cit.*, 45–46.

[380]*Op. cit.*, 53–57.

So justice is taken as an intrinsic good, and justice involves punishment being right under the right circumstances. The difference between this and explicitly affirming intrinsic-good retributivism is merely verbal.

Another question: why does it respect persons? This one can be answered: part of what we do when we punish with a retributive attitude in Duus-Otterström's sense is to affirm the punished wrongdoers as persons. This follows simply from the assumptions that we hold as part of retributivism.[381]

The real question is: why punish when you hold wrongdoers responsible? If it is all about our attitudes, could we not change our attitudes? Could we not show our respect in some other way? There is no conceptual connection between respect and punishment such that respecting requires punishing unless we already accept the premise that it does. The same goes for holding responsible and punishing.

It seems to be a very basic moral principle that we should not harm others. It may be taken that justification for punishment overrides this, but why? Should we take it to be so? If we accept the first principle, surely we should need very weighty reasons if we are to override it as greatly as we do in institutionally punishing people all the time, sometimes very severely. If your reaction to this is something like "Severe punishments are are not a problem if they are only for severe crimes," you are making the assumptions I am questioning. People are (in my experience) often ready to jump from the fact that someone has done something wrong to the normative conclusion that they may be harmed or should be punished, but when they do, they are definitely applying an unsaid retributivist premise.

If the goal in punishment were just to respect people as free persons, surely we should conclude that we have to do this in some other way to avoid the massive harm caused by punishment. In not really taking this into account, Duus-Otterström is again effectively assuming a picture where retribution intrinsically follows from something else.

[381]For Duus-Otterstöm's more detailed discussion on this, see 2007, pp. 174–196.

At one point, Duus-Otterström refers to James Rachels's necessary criteria for just punishment, which support retributivism over other criteria.[382] This is an interesting discussion in more than one way[383], but here, I focus only on Duus-Otterström's way of arguing for these, which contains a half-hidden appeal to intuition. Duus-Otterström states at one point that someone seeking to deny the premiss that these criteria apply must show what is wrong with them: it is not enough to point out that they are not argued for[384]. This is shifting the burden of proof to the wrong side. It is a perfectly sound counterargument to point out that the principles are not argued for. If one were allowed to pick principles without arguing for them, those principles could be anything whatsoever. Of course, in spite of putting his case like this, Duus-Otterström does not really implicitly apply the logic that anything would go. He continues to state that these principles are "very plausible" – effectively saying they are highly intuitive – and that applying the criteria as a test for justice only makes sense because of this[385]. Thus, his argument at this point amounts to an appeal to intuition, but he expresses it unclearly, even first verbally presenting a fallacious shifting of the burden of proof instead.

Duus-Otterström makes few explicit references to intuition using that word. Here is perhaps the clearest relevant case:

> Compatibilism ultimately relies on us accepting that questions of freedom and responsibility really concern the quality of determined choices and not whether the choices could have been different. Lots of intelligent people apparently accept just this. I have intuitions to the contrary. … I cannot see how ... punishing the normal shoplifter could be any less unjust than punishing the kleptomaniac, if [determinism]

[382] *Op. cit.*, 129–141.

[383] I discuss something touching on the same points in 8.2.2.

[384] Duus-Otterström 2007, p. 137.

[385] *Op. cit.*, 138.

is true.[386]

As comes up in chapter 3 of the present work, I am sure that there are different intuitions about freedom and determinism, and I do not essentially rely on any of them for making my own case. If compatibilist intuitions happen to be right in the light of well thought-out reasoning, bully for them, but they are not my proof or even (provided I am not illusioned about my own psychology here, which is always possible) starting point.[387] However, I purposefully refrained from saying much about responsibility when talking about freedom. I argued that we can be meaningfully free under determinism, but I will now say that does not actually prove that it makes sense to say we are responsible. That might seem tempting to conclude because of the automatic psychological connection between freedom and responsibility, and indeed I did state that the assumption is part of the intentional mode of thought (5.3), but it will need to be proven separately. At the same time, I hardly accept the other intuition-based thesis that we would need libertarian freedom to be responsible. At this point in the present work, some of its proponents have had a chance to argue for it, and yet, it has been given no better reasons than the case for compatibilist responsibility has been given by this point – even though I have not yet made that case for compatibilist responsibility.

Of course, Duus-Otterström already mentions reasons that could justify punishment without retributivist assumptions, particularly deterrence. We can follow this path, but where will it lead? If we reject assumptions amounting to saying that retribution is intrinsically good for some reason or other, are we led to abandon the idea of justice and retributivism in favour of utilitarianism?

---

[386] *Op. cit.*, 264.

[387] I would identify my main motivation for this study as starting from the realisation of the incompatibility of indeterminism and free will and the consequent wrongness of wanting to associate them.

6.3.3 Questioning Smilansky

Smilansky takes it as a given that justice and freedom in the full sense require full ultimate origination. Control compatibilism could not be a full answer because there are objective, non-negotiable requirements that it cannot fulfil. Hence, my refutation of his position is simple: he does not argue for this point, and I do not grant it. While aware of what he calls the "ultimate level", I do not grant that it has defining importance.

When Smilansky asserts the primitive importance of up-to-usness (on every level), he makes a sort of appeal to intuition, though it is not completely explicit.[388] It is almost as if he argues that we should not abandon the intuition because that would be bad. This is circular, and as we have reason to question the intuition, it will not do. However, it is easy enough to formulate this as an explicit appeal to intuition: this is a major intuition of ours, and that is enough of a reason to say that we must accept it as a moral truth. (Of course, this raises the usual questions of who "us" is. Smilansky's argument even involves saying that those of us who do not agree about this intuition are under illusion. While he can point to something that they are ignoring – at least provided their approach is not like mine here – it is still a case of being selective about what intuitions count.)

While we may be able to draw such a view from our intuitions, there is no reason as such to believe that there "exists" a morally important kind of freedom and responsibility that would require full ultimate origination. Remember[389] that Smilansky's own thesis is that the kind of freedom most worth wanting is a kind that is "incoherent", and therefore, "life is absurd." Do we really need to conclude the absurdity of life here? Would it not be better to conclude the misguidedness of wanting the kind of "libertarian free will" that he discusses – at least after realising

[388]One place where he explicitly says something – not much – about satisfying intuitions is in Smilansky 2011, p. 427.

[389]See 6.2 above.

these things about it?

Consider what the proposed state of affairs would be like: *there just is* an important kind of moral property that affects our judgements of the moral qualities of the world a great deal – which is never instantiated in the world and never could be. I do not think there can even be a detached moral fact not based on anything else.[390] It is stranger still if this moral fact cannot find proper referents in the world. It is even more strange than that if, going with the randomness argument all the way, this moral fact contains a demand for a self-contradictory property (of full ultimate origination).[391] If your starting assumptions lead to such conclusions no matter what the world is like, why conclude that the world is absurd for not matching your assumptions rather than that your assumptions are wrong? Would it not be less radical to allow critical questioning even strong intuitions – since those are really the only things holding this view in place – than adopt a view of the world as absurd?[392] Of course, from Smilansky's point of view, this is boldly facing (moral) reality, but the question is of how good the arguments are for the notion that this is how reality is, and then, following my thread of argumentation, it becomes a question of being willing to question your intuitions. Maybe what is *absurd* is instead the idea that "libertarian free will" is so important even after it is realised it is incoherent. Maybe it was an initially reasonable hypothesis that should be given up at this point.

Given that I do not accept intuitions as proof, what I just said is a persuasive argument, not proof. If we drop such arguments, we are just left with no reason to think the thesis about the importance of ultimate origination or libertarian free will would be true in the first place. On the other hand, if one still thinks intuitions can

[390]See my explanation of morality in Kokko 2018 for more explanation on how I think it really works. Also, see Nagel Nagel 2012, chapter 5 for a view also defending extremely detached moral facts, which to be clear is not something I agree with.

[391]Cf. Hurley 2000 for more discussion on the general question: whether responsibility is essentially incoherent and whether one can refer to impossible essences.

[392]Cf. Dennett 2015, p. 188. See also 4.3.4 in the present work.

be proof, from that point of view, my persuasive argument here may be seen as proof against the suggestion that intuitions about the importance of full ultimate origination are believable proof.

We can also question the notion that the world would be a better place if there were full ultimate origination, something that Smilansky believes. Well, perhaps *question* is not the right expression. It follows from his premisses but it does not follow without some of them. What I can really do here is present a perspective in which it can be seen as odd; as above, this illustrates my point, not proves it, since the "proof" would be based on intuition. So: Whether full ultimate origination exists or not, the world could be exactly the same otherwise, contain exactly the same amount of pain and pleasure, human achievement and opportunities, whatever you consider of value if it is something that affects the world in a concrete way – it could be exactly the same in these respects, but in one case, some opinions held by people as well as, notably, some suffering would be better things than in the other because they would be just. This would be the difference between a structurally just and morally superior world and a structurally unjust and absurd one.[393] (Cf. 2.5.)

Of course, the world *could not* really be like that, because full ultimate origination is impossible and no possible scenario can be derived from it, nor its consequences really determined. This is only a limited view dependent on ignoring some unavoidable facts; as it were pretending that we are living in Smilansky's illusion. However, that only makes it even harder to see what exactly we are supposed to long for in full ultimate origination. Apparently just being able to assert that things are just.[394]

---

[393]Smilansky states much the same thing but from a sympathetic point of view in Smilansky 2000, p. 49: "Libertarian free will, if it could exist, would have made this a *morally better world* in which, for instance, given that a large measure of punishment is bound to continue, such punishment could respond to people's ultimate desert."

[394]Of course, it is postulated by its proponents that this would be a *true* assertion, and its truth would be important, not its assertion. It is just that it has no truth-maker, and refers to nothing in the world, other than itself. Thus, it is fair to say it is really only that assertion, even if people with a certain kind of view would strongly feel that there is something more.

If you look at what Smilansky has written, it might seem that he has answered this objection. He addresses the general idea of the second main claim of the randomness argument that such free will is not worth wanting in a section in his book[395]. The point as he makes is that we can contrast the world as it would (impossibly) be with "libertarian free will" with this world and see that it would be a better place.[396] As he summarises it elsewhere: "The various things that free will could make possible, if it could exist, such as deep senses of desert, worth, and justification, *are* worth wanting."[397] However, this does not answer the objection at least as I am putting it here. Smilansky's counter is based on assuming that "libertarian free will" (full ultimate origination) *is* necessary for these valuable things. This itself is the basic assumption that is being questioned. To invalidate this defence, one needs only to clarify that the argument that an incoherent idea of freedom is not worth wanting also includes the idea that it is not justified to assume this incoherent freedom is a requirement for a deep sense of desert, worth or justification. Again, if it is an incoherent idea and cannot be realised in the world, why admit it as the foundation of things that are morally important to us? This just leads to postulating a "deep" sense of desert etc. that is really an empty sense.

Smilansky speaks of an "ultimate level" where no-one is free or responsible for anything. As I noted there, this parallels what I say in 5.3.3 about the mechanistic perspective in which we view everything as (event) causal. It is not the same thing,

---

[395]Smilansky 2000, 48–50.

[396]I will grant this here, and I am somewhat sympathetic to the idea of contrasting the kinds of worlds even if one is impossible, though I suspect it is not strictly speaking coherent to, as it were, examine a possible world that is not possible. The impossible world would have to both have some properties and not have them, and you could come to different conclusions about it based on which set of properties you functionally accepted. Thus, your conclusions about what would follow in the impossible world would be vulnerable to the objection that you might just as validly have come to the opposite conclusions by arbitrary choice.

[397]Smilansky 2011, p. 440, italics in original.

though. In Smilansky's ultimate perspective, there is no responsibility and every attribution of it is unjust because there are criteria that are not met. In my causal perspective, there is no such thing as responsibility or as justice – or injustice, for there are no criteria for them at all. Certainly it is possible to take the causal perspective and try to apply categories of freedom, responsibility and justice, and find that their criteria are never fulfilled. However, that can also be seen, not as a view revealing the deepest truth about freedom, responsibility and justice, but as a misapplication of those concepts within a perspective where they do not belong. That is how I see Smilansky's ultimate perspective: a confusion of categories. Now, I think it is a good observation that such categories cannot be applied in a causal perspective. Likewise, that trying to do so will result in their vanishing from view. It certainly helps understand the motives of incompatibilism. It also helps to understand the nature of human agency: that it is something that can be reduced, as it were to its components, and once you have done the reduction, you no longer see it.

A proper reduction or explanation of something that is real should leave one with two perspectives: one in which you cannot see the thing being reduced any more, since it has genuinely been broken down to its conceptual components; and another one where you still can see it in spite of the reduction, in spite of being aware of the other perspective. If you cannot see the thing at all after the reduction, it was not real in the first place; if you can see it in the *reduced perspective*, it has not really been explained since the reduction cannot do without referring to it (see 10.7 in the present work). This kind of proper explanation that neither makes the thing that is explained vanish nor preserves it unanalysed is what I try to do with freedom and responsibility in the present work. In Smilansky's analysis, they are seen as not having been real, and that is why the reduction destroys the unreduced perspective. He still has two perspectives, the control compatibilist and ultimate one, but for him, the latter proves the former to be incomplete.

I admit to a slight regret that I have to abandon a view as *interesting* as Smilansky's in favour of something more conventional – though that is not of course an argument for or against either his view or mine.

6.4 Questions going forward

We can observe that certain intuitions about responsibility, justice and punishment are commonplace. They are also followed to various degrees in different penal regimes. However, whether retributivist punishment is justified at all now emerges as a major question. We do harm to others, and a major defence of this, one that some thinkers claim must be the only one, is a mere appeal to intuition. Surely if there was ever a reason to question unquestioned intuitions, that is when they are used to justify institutionalised mass violence.

Punishment is just a side issue for the present work, however, if a major one. The central question it relates to is just how we should understand responsibility. We have seen that it is commonly assumed that responsibility is something that, among other things, requires freedom and autonomy and implies liability to consequences such as punishments. Duus-Otterström and Smilansky, for all their analysis, for much of the time treat responsibility as a given schema connecting these concepts and others. Responsibility relates to freedom, it relates to justice, it relates to autonomy, in certain ways, but ultimately, there is no explanation as to why it does. We gain some understanding as to what responsibility means for us as humans and how it functions in our minds by understanding it as a schema. Starting with the next chapter, I will begin answering deeper questions: why does such a schema exist, should we accept it and why, and how should we understand it?

7 Deconstructing Responsibility

> But whatever responsibility is, considered as a metaphysical state, unless can tie it to some recognizable social desideratum, it will have no rational claim on our esteem. Why would anyone care whether or not he had property of responsibility (for some particular deed, or in general)?
> –Daniel C. Dennett[398]

> Thinking through, critically and carefully, what most people take for granted is, I believe, the chief task of philosophy, and it is this task that makes philosophy a worthwhile activity.
> –Peter Singer[399]

This chapter continues the project begun at the end of the previous chapter of questioning common assumptions related to responsibility, justice and punishments. First, I clarify things by taking a look at different things that can be meant by "responsibility". Then I catalogue the assumptions we have and reduce them to their bare bones about connections between such things as free will, moral responsibility, and punishment. I continue by questioning these assumptions, especially about the justification of punishment as an intuitive requirement. I do this both via direct argumentation elaborating on what I did in the last chapter in the last chapter and via different approaches related to modern psychological theories that suggest there is something wrong with retributivism. Thus responsibility is "deconstructed" both by

[398]Dennett 2015, p. 178.

[399]Singer n.d., p. 6.

taking it down to its component parts and then, tentatively, by revealing what is possibly hiding behind it.

7.1 Kinds of responsibility

Though I have mostly been speaking of responsibility as moral responsibility and of moral responsibility as if it is one thing, there are many different senses of responsibility – different senses of moral responsibility, senses that are not moral, and senses that can be either moral or otherwise. All of these may have something in common, but they must not be confused. In the following, I introduce typology of kinds of responsibility that serves to make clear which kinds of responsibility I am talking about at each point and which kinds – not all moral – I can draw conclusions about based on my theory.

The typology I employ here is in large part based on Nicole Vincent's taxonomy of responsibility concepts[400], which is a comprehensive and thoughtful examination of the concept, as well as informed directly by H. L. A. Hart's discussion that Vincent is also drawing from.[401]

The first form of "responsibility" to be noted here is **causal responsibility**.[402] To be causally responsible for *X* is basically to be the cause of *X*, and thus, either agents or non-agent things can be causally responsible for something. Questions of agents being causally responsible or not for things have come up in the free will debate repeatedly, but beyond how I have already discussed those questions and how I will later,[403] they are not particularly relevant here. Causal responsibility is required

[400]Vincent 2011.
One kind type of responsibility recognised by Vincent that I will not be discussing at all is *virtue responsibility*, which could also be called "responsibleness". This concept is about being a responsible sort of person (Vincent 2011, p. 16).

[401]Hart 2008, pp. 211–230.

[402]Vincent 2011, p. 18.

[403]See 3.4.1, 3.4.2, 9.6.1.

for outcome responsibility[404], but that is a fairly trivial observation in the present context.

As suggested by the previous discussion, **capacity responsibility** is about whether the agent has the capacities required to take or not take the action they would be held responsible for, either mentally or physically.[405] Examples of lacking capacity responsibility include someone who is too psychotic to know what is real, a child with undeveloped mental capacities, someone whose body is moving convulsively, or even someone who simply lacks the physical strength to do what they might have been responsible to do. Even basic attributions of causal responsibility in the case of persons require a certain minimum amount of capacity responsibility: if someone's physical body causes something to happen in a way that does not involve the person taking *an action* in the normal sense, that is not counted as being even causally caused by the person[406], at least as far as causal responsibility as a requirement for other kinds of responsibility goes. Capacity responsibility is also a requirement for *role responsibility*.

Role responsibility (taken broadly) is basically who is being held responsible to do what, in advance.[407] Besides of more obvious roles like a ship's captain having certain responsibilities, it can be taken to encompass (as I do here) everything about what is expected of people in their respective situations.[408] In the discussions before this section, insofar as it has made a difference, I have been assuming that responsibility is something backwards-looking, and about actions. Role

---

[404] *Op. cit.*, p. 20.

[405] *Op. cit.*, p. 18.

[406] *Op. cit.*, p. 1.

[407] *Op. cit.*, p. 17.

[408] *Op. cit.*, p. 20-21.

responsibility is basically forward-looking, amounting to the statement that the person responsible *should* act so as to take responsibility for the thing they are responsible for. It can also be backward-looking in that we can judge people's actions in a certain way *because* the person was responsible for a particular thing. Capacity responsibility is required for this kind of responsibility in that people are not held responsible to do what they cannot do due to their capacities[409].

Outcome responsibility refers to when an event is held as attributable to an agent's actions.[410] It requires both causal responsibility – the agent must have played a causal role in the event happening, of course – and role responsibility, since if the agent is not being held role-responsible, that definitionally means they are not being asked to take responsibility or behave in a certain way in that context. It follows from this that if two people are both being considered as possibly responsible for a bad event, then who is responsible is judged by who was doing something wrong in the first place. In an example, person *A* shooting person *B* to death with no justification would be responsible for *B* dying because *A* was doing something *A* should not do, whereas if *A's* shooting *B* is seen as justified on the basis that *B* attacked *A* and *A* was acting in self-defence, this can be explained by appealing to how *B* was already doing something he should not.[411] Notice how "role" responsibility gets extremely broad here, in such examples arguably encompassing one's role as a fellow human being or agent in general, with such agents having a *prima facie* responsibility[412] not to harm each other. I will continue to use it in such a broad sense.

The last item from Vincent's list to be discussed here is **liability**

[409]Vincent 2011, pp. 20–21.

[410]*Op. cit.*, p. 17.

[411]*Op. cit.*, p. 20.

[412]We could speak of duties here in place of responsibilities. Since we *can* also speak of being responsible instead of having a duty, I will stick to the terminology of responsibility for simplicity. (Cf. *Op. cit.*, p. 17, footnote 5.)

responsibility. This means someone being held responsible for what happened in the sense of being liable to consequences such as punishment or being held financially responsible.[413] A big part of the relevance of outcome responsibility is that it is a requirement for liability responsibility.

The kind of responsibility I am mostly concerned with in this thesis is moral responsibility. What is the relationship between the above mentioned types of responsibility and moral responsibility? Causal responsibility in itself is simply not moral, though it has relevance to moral responsibility in the sense that it is required for other kinds of responsibility that may be moral. The same thing can be said about capacity responsibility, though it has a more morality-adjacent flavour as a concept.

Meanwhile, role responsibility, outcome responsibility and liability responsibility can all be moral, though they do not need to be. We can say that you are morally responsible to do something or not given your position (role responsibility), morally responsible for an outcome, and morally liable to suffer or enjoy the consequences of your actions (liability). However, we can analogously say that the law says a political candidate is responsible to disclose where they got their campaign funding; this would be role responsibility that is legal responsibility. We could also say that a university student is being held responsible for the advancement of their own studies in the sense that their teachers are not holding their hand and checking that they are completing enough courses. This is role responsibility with no other obvious descriptor, and it is an interesting case that will be studied a little more in 9.3 and 9.5, since even though it is not moral responsibility, it turns out to be explained well by the theory I build for explaining moral responsibility. Similarly, one could morally blameworthy or praiseworthy for an outcome, or responsible for the outcome in the eyes of the law, but also in some other sense, such as when considering who really contributed what to a project. This could carry on to liability responsibility in the person morally deserving to be treated in some way, in being

[413] *Op. cit.*, p. 18.

legally liable to punishment, or deserving more or less compensation for their role in the project.

Thus, for role responsibility, outcome responsibility, and liability responsibility, we have at least a moral version, a legal version, and a third, partly moralistic and partly pragmatic third version without a particular name. Because these all share in common the features belonging to role, outcome, and liability responsibility, my discussion of moral responsibility is going to have some bearing on the other kinds as well.

7.2 The schematic features of freedom and moral responsibility

If we generally take some things for granted about the relationships of concepts like free will and responsibility, and if I want to examine these assumptions here, it is best for me to spell them out as clearly as possible. Below, I will do so with what I see as the most central such assumptions. There is an obvious partial overlap with the analysis borrowed from Vincent, which I will not spell out here, though I will discuss how both of these are covered by my own theory in section 9.4.

I call these "schematic features" because I see them as representing a schema relating some concepts and conditions to each other, especially the central concepts of "free will", "(moral) responsibility", and "being able to do otherwise" as well as the concepts of *praise*, *blame*, *reward* and *punishment*.[414] We have already seen that these concepts can be given different meanings but still employed at least roughly the same way in the schema, for example if a compatibilist and libertarian both think it is the kind of free will that they endorse that implies moral responsibility. I will

[414]I use first quotation marks and then italicisation in accordance with the practice I described in 1.5: The first three mentioned "concepts" in the sense they are used in this sentence are not attached to specified meanings (so they are "words" rather than "concepts"), where as the last four I take to have uncontroversial (broad) meanings I am employing in their normal sense here. Also see the definition of *punishment* in 6.1.1.

also make use of this schema to show how my theory explains and justifies it later, especially 9.4.

I am suggesting that these schematic features reflect common intuitions. I will give some references here as examples of authors endorsing them, though these are just a tiny sample per each. Chapter 6 showed examples of how these sorts of connections can be taken as intuitively given. Then again, the same premisses may be accepted based on other kinds of arguments, which is something I discuss in chapters 8 and 9. They may also be accepted because they are seen as generally accepted. It would take a major study in its own right to find the most representative examples of thinkers endorsing these ideas, let alone to list who endorses them based on intuition or other reasons. Thus, the list of references here is only an unstructured sampling to show that these ideas indeed have people endorsing them (certainly not all on intuition). The examples are taken from quite various and different sources, which at least suggests the widespread acceptance of these assumptions.

1. **Having free will requires the ability to do otherwise.**[415]
2. **Being responsible requires free will.**[416]
3. **Being responsible requires the ability to do otherwise.**[417]
4. **Being responsible for a good or bad act implies being liable to good or bad treatment: reward, praise, punishment, censure.**[418]

[415]E.g. Dennett 2015, p. 143, Tononi 2013, p. 174, Harris 2013, p. 10, Reid 2010, p. 223. See also 2.4.2 in the present work for discussion of this as a general principle.

[416]E.g. Honderich 2002, p. 110, Schlossberger 1992, p. 10, Willmot 2016, p. 8, Slattery 2014, p. 232, Hart 2008, p. 28. This is of course a major premise for both of the authors discussed in chapter 6 in the present work.

[417]E.g. Dennett 2015, p. 143, Schlossberger 1992, p. 3, List 2019b, p. 23. Of course, this is denied in specific positions like semicompatibilism (Fischer & Ravizza 2000), but that is treated as a substantial claim made against the usual assumptions.

[418]E.g. Strawson n.d., Schlossberger 1992, p. 1, Dennett 2015, p. 172, Reid 2010, p. 158, Honderich, 2002, pp. 134, 141. Of course, also see from earlier in this work chapter 6 again, and 7.1 for liability.

5. **Being responsible implies the absence of excusing conditions, which include at least: ignorance (within limits), lack of understanding, coercion, absolutely or relative inability to do the right thing, belonging to an excused group or category (non-humans, children, the "insane"), and being determined in an unusual way such as an illness.**[419]

Now, it is not that I do not personally share these intuitions. I share them at least as strongly as I do the ones concerning free will and determinism or alternative possibilities. I am easily capable of practising the kind of reasoning and/or intuiting where I consider, for example, whether someone could be responsible if they had no option to do otherwise, and coming up with the automatic answer that this cannot be so. I even find plenty of retributivist feeling and automatic reasoning in myself. It is simply my considered opinion that this is not enough for a philosopher or any moral reasoner to do – and I do not consider appeal to my personal intuitions any more valid than appeal to those of others[420], nor do I even find it hugely more compelling, though of course I find it *somewhat more* compelling almost by definition by virtue of it being my own intuition.

As a result of all this, I need to ask whether these premisses often derived from intuitions can actually be defended as valid.

---

[419]The specifics here go into so much detail that many authors do not explicitly discuss the specific ones unless they are discussing this topic specifically, like Hart 2008, p. 28 and chapter II, so I do not have a sampling from different kinds of sources to offer here – except for the "sickness" kind of examples: e.g. Harris 2013, pp. 8–9 and 52–53, Willmot 2016, ch. 2 and 4, Schlossberger 1992, p. 73, Reid 2010, pp. 199 and 223 – and about children: e.g. Harris, 2013 pp. 51–52, Schlossberger 1992, p. 73, Reid 2010, p. 223, Hart 2008, p. 19.

[420]On appealing to intuitions, see 6.3.1 in the present work.

7.3 Responsibility and differential moral treatment

In terms of implications for concrete ethical choices, one of the most significant conclusions from these notions of responsibility is what follows from liability responsibility: how it is thought to be right to treat other people based on what they are responsible for. The usual list is that responsibility is what makes one liable to praise and blame, reward and punishment. If someone has done something good, they deserve positive things: praise or rewards. If someone is responsible for having done something bad, they deserve negative things done to them: blame or even punishment.

Outside of a context of desert, the ethical imperative is usually to treat other people well. Responsibility, in these common intuitions, is thought to remake the rules. If a person "deserves" some particular kind of treatment, this statement hardly has any other content than that it is right for others to treat that person that way – either that it is a good thing, or that there is even a duty to do that. Thus, doing harm may become the right thing, doing good can become more right than usually, and doing good can even become wrong insofar as someone does not deserve something.

The concepts of praise, blame, reward and punishment depend in their meaning on the discourse of desert and liability. We can always try to look at them ethically outside it too. Aside from that context, to praise someone is generally an act of being kind to them and to blame them is generally to be hurtful towards them. A reward is also a kindness, and it is a fairly obvious direction to start thinking about rewards in terms of their effects on reinforcing or discouraging certain kinds of behaviours.[421]

The upshot is that ideas about responsibility are thought to ethically justify treating people differently from each other. This is why their justification is a question that matters ethically. When responsibility is treated as ethically primitive

[421]For all of these, we can also raise considerations of *fairness*, but that concept in this context overlaps heavily with the concept of desert and is just as intuitive, and I will not consider it as a separate category here.

based on intuitions, those intuitions are taken to justify treating people differently in an ethically relevant manner, and this includes the right or obligation to do harm to others.

As discussed in chapter 6, this moral claim is the heaviest when applied to punishment. Punishment is straightforwardly inflicting harm in principle. This makes punishment the most morally important implication of liability. I will now continue by treating punishment as both the biggest question and a proxy for all the questions about treating people differently based on notions of desert – what applies to it applies to praise, blame and reward, though (usually) less importantly. The core question is whether it is right to treat people differently based on notions of responsibility and desert, especially on the basis of intuitions alone.

7.4 Against retributivism

Retributivism is the idea that punishment for the right reasons is a morally good thing. We can exclude ideas in which it is thought to be good to some consequentialist end, and given that, we can treat retributivism as being the same as intrinsic-good retributivism (see 6.3.3). It does not matter if other concepts are inserted, such as that punishment is good not in itself but because of desert, since the result will be the same for individual moral judgements about the rightness of acts. Retributivism amounts to holding that if someone has done something bad and there are no mitigating conditions, then it just is right to harm that person. This sits oddly with the rest of morality. In this section, I argue why it cannot be defended as an intrinsic value.

The contradiction in the idea of morality involving both compassion and the principle of not doing harm on one hand, and retributivism on the other, shows up in ultimate form in Christian ideas about God's supreme goodness. The idea of a God who is loving but also just – with justice seen as involving the intentional infliction of harm, since it involves punishment – is an uneasy combination. The ultimate punishment is the idea of those humans who are not saved spending an eternity in hell, and throughout the history of the idea of hell, the enormity of this idea has

repeatedly risen as a problem for thinkers considering the matter[422]. Yet, this is only an exaggerated analogue of what goes on in mortal moral thinking all the time.

7.4.1 Intrinsic goods

> Anybody, it seems, can declare that something is an intrinsic good. I can now say that it is an intrinsic good that the guilty should be reformed, or better that the only intrinsic good is the prevention of suffering.
> –Ted Honderich[423]

If some things are morally valuable only instrumentally, there must be something that they are instruments towards: some things that have intrinsic value. The previous chapter brought up the question of intrinsic-good retributivism, or broadly speaking just retributivism: the idea that, once there is something to punish, punishment is itself an intrinsic good. Both Duus-Otterström and Smilansky even took the stance that to punish for the sake of what consequences are expected from doing so would be morally objectionable.

As I already began to argue at the end of the last chapter, this retributivist moral high ground is questionable when considered from the opposite direction. It is not merely saying that punishment in the service of the greater good in terms of consequences is wrong, but also that punishment is still good in spite of its negative consequences. The normal moral imperative to do no harm is overridden by another imperative to do harm in the right way. Thus, the reverse of normal consequentialism or utilitarianism comes into force: in the right kind of circumstances, aiming for good consequences is not permitted, but aiming for harm to be inflicted is right. The half of this equation that I want to focus here is the second one: it becomes right to do

[422]See Kuula 2006.

[423]Honderich 2002, p. 138.

harm. This is a very major moral claim, and it is typically supposed to be accepted on intuition. I have already argued against this in the previous chapter, and this chapter will deconstruct the idea even further.

The basic argument for utilitarianism in general and especially for doing no harm specifically is simple and grounded in external reality. Accepting its basics does not rely on accepting a wholly utilitarian theory, either. In a brief formulation of my own:[424] Perhaps the only thing that we are compelled, by the nature of that thing, to regard as good in itself, is positive feeling – pleasure, enjoyment, and so on. Correspondingly, the only thing we are almost universally compelled to think of as negative when experiencing it is pain, suffering and so on. These positive and negative values are imposed on us almost universally and unavoidably. Though it is not possible to logically draw normative conclusions from any facts about the world, even these, we are practically compelled to take them into account. Going against them is going to almost automatically lead to something that we will be compelled to judge as more negative than positive. Therefore, these things that I will call pleasure and pain for simplicity are at least one basis of value that we are extremely strongly and probably inevitably compelled by practical reasons to hold as important values.

This is only a brief argument, though I think it is objectively extremely powerful. There is no space to argue for the positive moral value of pleasure and the negative moral value of pain in more detail.[425] This, of course, is ironic considering that I argue here that this is a much stronger, more reasoned and better able to stand up to reason, and more significant basis of morality than (intrinsic-value) retributivism, and that is because retributivism is not given sufficient grounds. All I

[424]This parallels some earlier arguments for utilitarianism, of course, such as Jeremy Bentham's, but differs variously from them, a question that there is no space to explore here; for a quick overview, see Driver 2022.

[425]Some more detail is given in Kokko, 2018. This piece also addresses the question of how intrinsic goods can be "derived" from empirical factors.

can add in this space is that the positive value of pleasure and the negative value of pain, whether in the garb of consequentialism or something else like deontology, is also a basis of value most people seem to agree about even in cases when they also agree with retributivism. Even intuitively, it is at least as strong as retributivism. If someone really wants to contest this basic value altogether, that will need a separate discussion to address it.

Given that we already have the important moral status of it being wrong to inflict harm on others, then what I argued at the end of the last chapter holds, and we cannot just accept an intrinsic positive value of harming others for the sake of harming them, based on intuitions. This is my main argument against retributivism, and I hold it to be sufficient to prove its conclusion. I will now continue with other arguments that do not have the power to *prove* anything but which do *support* the same conclusion against retributivism. I will spend more time on those arguments than this more important one because not only did I already start making this one in the last chapter, but it is also very simple, and if it needs elaboration, that is more to persuade people of it or illuminate it than to state it – and the following suggestive arguments are also persuasive and illustrative arguments for the general point that intuitions cannot and do not prove the correctness of retributivism.

7.4.2 On the psychology of retributivism

If we commonly have retributivist intuitions (which seems extremely likely), our having them is a psychological fact about us. If we view it as such, we can ask psychological questions about it. Where do retributivist intuitions come from? How do they function psychologically? While it is hardly possible for me to prove anything about these questions with respect to my philosophical opponents on this issue, a look at what psychology has to say about such intuitions in general may give some useful perspective. This, too, is a large topic that I can only glance at briefly here.

It should be noted that what I aim to do is exactly what I stated above: to give some perspective on how to think about the question. These observations do not

outright prove anything about the philosophical and ethical questions. They might, however, give us tools for relating to our intuitions.[426]

Firstly, is there an explanation for where retributive intuitions come from? A possible one is found in a mixture of evolutionary psychology and game theory.[427] Put shortly, the idea is as follows: In interacting with other people, we are constantly faced with the choice of whether to play fairly, do our own part, and not violate the other person's interest – or to do the opposite. For short, the first option is usually called cooperating and the second defecting.[428] A kind of paradoxical situation that arises is that everyone would benefit more from cooperation overall, but it is generally possible to gain a short-term advantage by defecting in a given situation. This raises the question of what kind of a strategy would be most beneficial, or even sustainable. Pure cooperators would fare better as a group, but a defector playing against cooperators will always win at their expense. In the sense of evolution, this becomes a question of which gene variants can outcompete others in a population by inclining their bearers to behave in ways that bring benefit to their bearers. If we only consider cooperators and defectors, the population would come to be dominated by defectors, to everyone's detriment, leading to some version of a war of all against all as nobody could trust others not to try to exploit them. A group of cooperators would do better than a group of defectors, even outcompete such a group, but since natural selection is not truly teleological, it cannot "plan" to create a group of cooperators when defectors outcompete them within the group.

---

[426]Arguments trying to disprove an idea based on its origin are known among other things as *genetic arguments* or (more commonly) *debunking arguments*. (See e.g. Korman 2019.) As I state, I do not go as far here as to claim that my genealogical suggestions outright disprove the idea.

[427]This evolutionary story is summarised in e.g. Dawkins 2016, chapter 12, Sapolsky 2018, pp. 342–354, and Pinker, 2011 pp. 532–536.

[428]This situation is usually called a *prisoner's dilemma* situation; for brevity and variety, I will avoid the usual overused (though undeniably handy) example of prisoners deciding whether to testify against each other.

There is, however, a still fairly simple strategy that outcompetes defectors by not being vulnerable to it while also being able to benefit from cooperation. Dubbed "tit-for-tat", this strategy starts off by cooperating but then mirrors whatever move the other party did last round. Thus, if met with a cooperator or another tit-for-tat user, it will be nice and reap the benefits of cooperation, but met with a defector, it will turn mean and not let the defector benefit from the exchange.

If such a strategy were to be genetically enforced, what would it look like? It could, and quite plausibly does, take the form of becoming angry and wanting to inflict punishment on those who do something wrong. Elaborate theories aside, this is the retributive impulse.[429]

If this is the original source of retributivist impulses, one that existed before all the theories, such a causal explanation and origin for them may not feel like it supports their revealing fundamental moral truths. That would be an odd coincidence, at least.[430] Further, note how this very evolutionary-psychological origin story itself is based on a kind of consequentialist and deterrentist logic: it happened because it allowed for better results by making crime not pay. This, again,

[429]Incidentally, rejecting retributivism is not proof that the one rejecting it does not have this genetic tendency. I clearly have such intuitions, wherever they come from; I sometimes feel moral anger and a desire that someone be punished. I just do not believe them to be morally right as such – though the part that *only* the appropriately guilty should be punished I do consider morally good; see 8.2.2 for some reasons for this – and, though it is perfectly possible they would cloud my judgement in some case, I reject them in principle based on both reason and on opposite feelings like compassion that I endorse and cultivate more. (Though I have not formed any theory based on this, I have toyed with the idea that, in an ideal world, the only pain that should be inflicted on someone for their wrongdoing would be a pain from realising what they have done that would lead to their character being improved. I have explored this idea in a short story (Kokko, 2022). Still, the hypothetical infliction of such a pain perhaps satisfies my retributive impulses as well as some nobler ones.)

[430]As mentioned in 6.3.1 (p.159), Thomas Nagel has argued against the possible causal evolutionary origins of our knowledge of objective values on a similar basis – that it would be an impossible coincidence – though note that I mention the argument there as an example of a bad appeal to intuition. What I agree with him about, in this context, is that you cannot expect correct knowledge of values to evolve for reasons that are themselves unrelated to the correctness of the values.

does not prove anything, but it puts the thought that the intuitions themselves would prove of something that is the opposite of consequentialism in an odd light.

So much for the possible evolutionary origins of retributivism. Regardless of whether that explanation is correct, what is the (social) psychology of retribution like in the present? In a study of beliefs in supernatural devil-like figures throughout history, Gerald Messadié writes:

> All I mean to say is that it is demonic to believe in the Devil, not the other way around, as the Jesuits tried to teach me when they said the Devil's greatest trick is to make one believe he doesn't exist.[431]

Messadié's point is that the belief that your enemy is associated with the devil is itself a root of evil because of how it makes you treat your enemy. This point goes beyond belief in literal supernatural devils, though.[432]

Roy F. Baumeister's book *Evil: Inside Human Violence and Cruelty*[433] is a thorough study on different aspects on humans causing harm to each other, particularly the reasons. I focus here on one side of the question: how we are often blind to what is really happening when both ourselves and others cause harm. Shortly put, we are biased to think as if only especially wicked people can commit evil deeds, and this has some very worrying consequences.

The idea that evil can be perpetrated by ordinary persons, not only moral monsters, is popularly known as "the banality of evil" and attributed to Hannah Arendt and her book *Eichmann in Jerusalem: A Report on the Banality of Evil*[434]. In

[431]Messadié 1996, p. 10.

[432]Cf. Messadié 1996, pp. 6–7.

[433]Baumeister 1996.

[434]Arendt 1994.

reality, that book does not clearly discuss the idea,[435] and in reference to Eichmann, there is a case to be made that this meaning of "the banality of evil" is a misinterpretation of what she meant.[436] Baumeister's ideas correspond to it much more clearly while giving it additional depth explaining how it really works and why it surprises us.

A more familiar starting point that acts as as it were as the first step towards the biases Baumeister reveals is the *fundamental attribution error*. This is a psychological phenomenon where people tend to emphasise the circumstances much more when interpreting the reasons of their own behaviour and intent and character much more when interpreting that of others. This can lead us to more easily excuse ourselves from possible blame for our actions but to conversely judge others more easily.[437]

Baumeister presents evidence that we have biases that go much further still in this direction. Speaking of actions that harm someone else (Baumeister's operative definition of "evil"), we tend to rationalise such actions when we (or those we sympathise with)[438] do them, believing that we had little choice and that the harm

[435]Though it would take too much effort to ascertain this reliably, if my memory is correct, the only time the phrase "banality of evil" is used in the book besides the subtitle is as the last words of the final chapter (p. 252, the final words of the book not counting the epilogue and postscript).

[436]One version of such a case is made in Ushpiz 2016. It is not obvious to me whether this interpretation is correct either, but it seems worth considering.
My own and possibly shallow take is that while Arendt brings up that Eichmann seemed surprisingly un-monstrous for one of the architects of the Holocaust, she does not present him as being just a normal person either, but rather exceptionally unthinking and manipulable.
Baumeister mentions "banality of evil" in connection with Arendt at 1996, p. 379, characterising it somewhat vaguely such that it could be understood either in the sense I am criticising or as compatible with the sense that I just described as my own impression of what Arendt meant.

[437]Note the connection to the three modes of thinking described in 5.3 in the present work.

[438]In this discussion, I refer to the generic person experiencing these biases with the pronoun "we" to, hopefully, communicate a sense of how these are biases that affect everyone, and should importantly be considered in our own case, regardless of who

caused was not all that great. Conversely, and perhaps even more alarmingly, when someone inflicts harm on us (or those we sympathise with), we have a tendency to jump to thinking the act was very much unjustified and probably done out of malice, as well as exaggerate the harm. Even just looking at an imaginary scenario from a different participant character's point of view evokes these biases. Further, there is a bias to assume that the victim of that evil wrongdoer is pure and innocent, even though realistically this may either be the case or not.[439] These biases are also subtle but pervasive: they still affect people who would not explicitly affirm, say, that bad things are only done by bad people for bad reasons.[440] That is why I wrote above that we are biased to think *as if* only especially wicked people can do bad things, not that we are biased to *think* so.

These biases and their interaction with culture lead to what may be called *the myth of pure evil*:[441] the idea that bad things are only done by especially bad people with evil motives. This idea leads not only to glibly condemning the wrongdoers, but arguably also to the converse and just as dangerous illusion that *since we are good people, the things that we want to do cannot be bad*.[442] Thus, people doing harm to

---

we are. I also mean it to imply that we can apply this logic to someone else we sympathise with the same way as to ourselves, though I will not keep separately stating that throughout.

[439]Baumeister 1996, chapter 2.

[440]As an informal observation of my own, it seems that noticing that humans can commonly do bad things may not topple this logic but lead to a reaction that is in accordance with it: instead of coming to a compassionate realisation that wrongdoers need not be monsters but can be fellow travellers who stumbled in their search for what they see as good, or simply did not realise they were in a moral situation, it can lead to a misanthropic sense that humans in general are internally wicked and condemnable. In other words, the logic apparently goes: only bad people can do bad things, and everyone can do bad things, so everyone is bad.

[441]I take the term from Baumeister (1996, see especially chapter 3), but I use it more loosely and generally here than he does. In pp. 375–376, he characterises the myth in such a way that it would fit my use as well.

[442]*Op. cit.,* pp. 95–96.

other people still tend to see themselves as being in the right, and it is not too natural for them to stop to consider whether they might be. Even violent criminals whose deeds seem obviously wrong from the outside can rationalise and belittle what they have done – and even feel like *they* are the victim.[443] Further, moral conviction can easily even be the cause of egregious wrongdoing, such as in the case of a terrorist who sees themselves as fighting for important ideals.[444, 445]

It seems to me that there is a psychological connection or at least affinity between the view of evildoers as "containing evil" within themselves and intrinsic-good retributivism. Admittedly, the connection is hard to pin down, let alone prove, and my reasons for holding it come as much from introspection as from observing others' reactions. I will simply sketch it out briefly without making strong claims about it, since I think it is worth considering. The idea is that the evil that seems to taint someone's soul feels as though it is part of their essence, like a metaphysical aspect even though it is really in our own emotions, and that makes the person either temporarily tainted or even permanently essentially different from those we count as people who should be helped and not harmed. This, in turn, may be part of what gives us the kind of biases we do have towards just wanting to punish for its own sake. I would say it is worth for each of us to consider whether something like this is going on within us when we feel moral anger and retributive inclinations.[446]

---

[443] *Op. cit.,* pp. 47–52.

[444] *Op. cit.,* chapter 6. See also pp. 27–29.

[445] Think of Luke Skywalker in the original first *Star Wars* movie blowing up the Death Star: certainly, he killed presumably thousands of people onboard, but it was a planet-destroying weapon being employed for indiscriminate mass murder for the purpose of government terror, and it was just about to destroy the Rebel Alliance base at that very moment. Never has mass killing been so necessary and justified. Apparently, terrorists often see themselves as the heroes of such a story.

[446] For more on the not so admirable psychology of retaliation, both evolutionary and social psychological, see Pinker 2011, pp. 529–541.

It is certainly impossible to make universal statements about the myriad and complex motives of all violence, doing harm, and wrongdoing. Still, insofar as the above results are even roughly correct, we can derive from them a certain caricature – opposite of the myth of pure evil – that cases in the real world surprisingly often follow in part or even wholly. The caricature is this: we harm others while being convinced that we are acting out of understandable reasons, our actions are morally justified, and that what harm we do does not matter much, unless it matters positively. Meanwhile, when someone harms us, we assume it is because they are an evil monster out to do harm without good reason, and that they should therefore be righteously harmed, which is a completely different thing than what they did. Our brains and to some extent cultures are built to trick us into thinking this way.

So, it seems that what Messadié almost said in his above quote is correct: *The devil's greatest trick is to make us believe he exists.* If we imagine the devil as a pure evil force whose goal is to promote evil deeds in the world, then he could have hardly done better than to make us inclined to believe that malignity (supernatural or not) is a thing that resides in other people and righteously justifies our harming them. This myth of pure evil has led and does lead to people harming others all the time. It does the devil's work all the better in a world in which he does not exist and the potential for evil is something we all share instead.

To say an obvious thing that needs to be said, the philosopher's (perhaps) cool intuitions are not identical to these psychological phenomena roughly applying to people in general.[447] Whether they have the same origins and to what extent they may involve the same kinds of psychological phenomena is not proven here, though the hypothesis that they are a version of the same thing is plausible. (If they are not, it makes it complicated to claim that these intuitions are shared between the philosopher and most people.) Just because ordinary people are often driven by moral anger involving a certain degree of unconscious blindness does not necessarily

[447]Cf. e.g. Duus-Otterström 2007, pp. 96–99.

mean the same processes operate in, say, Göran Duus-Otterström calmly judging that to respect another person as an agent involves the obligation to purposefully harm that person afterwards if they committed wrongdoing[448] (see 6.1.2). Nevertheless, the possibility of some of the same processes or qualities being involved in the philosopher's reflections should certainly invite caution before going all in on believing the intuitions in such an important matter with such weighty consequences.

I wrote at the beginning of this section that the empirical observations in it are no more proof of philosophical propositions than intuitions are, but there is a flipside to this. If you *are* willing to entertain the idea that intuitions may offer proof of moral truths, the observations here are (inconclusive but arguably strong) evidence against the idea that retributive intuitions do offer such proof. After all, if we are to decide whether these intuitions are ones that are proof of objective moral values, then to answer that question, it helps to understand their origins and nature. If it turns out that a possible origin for intrinsic-good-retributivist intuitions is blind evolution acting in a manner analogous to consequentialist deterrence, and that the psychology of retributive impulses in the real world is typically delusional about the facts and its effects often destructive, this speaks against the likelihood of retributive intuitions being ones that can be relied on to tell us objective truths.

If you are not willing to consider the possibility that your retributive intuitions could be wrong, you might still want to meditate on the sources that have been presented in this section.

7.5 Conclusion to chapter 7

In this chapter, I have shown that there exists a set of interrelations between concepts such as free will, the ability to do otherwise, moral responsibility, and punishment – and that these interrelations, while very often accepted, all seem to rely on intuitions whose strength and weakness both lies in how naturally they are taken as

[448] *Op. cit*, chapter 6.

unquestioned and given. I have argued that it is not acceptable to make these philosophical and moral inferences, especially the justification of infliction of harm in punishment, based on mere strong intuitions. I also showed that some of these intuitions may have an evolutionary origin, and that they are at least related to psychological tendencies that give us a distorted view of the world, further calling them into question.

The whole conceptual field of moral responsibility, including the associated questions of freedom, seems to be defined by intuitions and beliefs that are *just there*, so ingrained that people usually never think of questioning them. They are also so ingrained that questioning them often seems impossible to people; if they try to imagine an alternative, they seem to run into a contradiction, as with the case of vanishing responsibility (1.1.3). It is like a circular house of cards built upon itself. It would be poor philosophy to let this stand without questioning, and this chapter showed that in some sense, questioning brought it all down.

Yet, I do not propose abandoning all these notions in all their forms. Perhaps the reason we have such intuitions that may seem to be about primitive moral values is that they really have a place in a moral system that is based on something else than every component having moral value as a brute fact. Maybe, if we can show this, we can also show that the system does not need to be argued for based only on intuitions.

If what I just said were really a guess at a possibility, it would be a bit of a leap, but it is actually hindsight, because that is just what I aim to show in the next chapter: we can reconstruct our notions of responsibility based on more fundamental moral values, and what we end up with is remarkably similar to what we had when just going by intuitions. I will also endeavour to show that, contrary to how it may seem, this makes the system of moral responsibility more important than if it had its own primitive value.

8 Reconstructing Responsibility

In the previous chapters, I argued that we cannot accept the moral ideas related to responsibility and especially the differential treatment of people it implies as intrinsic goods based on intuitions. As exemplified by Duus-Otterström and Smilansky in chapter 6, some people think that the only morally right concepts of responsibility and desert are ones based on intrinsic value like this, and no consequentialist instrumental justifications for things like punishment can be valid. Yet, as also seen with those authors, such a stance itself needs to be based on intuition. As we also saw in chapter 6, there are also many possible alternative rationales for using such concepts, such as a hoped-for deterrent effect.

In this chapter, I show how a model of responsibility and desert – and again, especially punishment – based on deterrence can complement what I have said about free will very well. Specifically, I show that our intuitions and practices of how responsibility and desert can match such a rationale. Given the schematic features of freedom and responsibility I identified earlier, as well as more specific points about out practices and beliefs as they exist, the kind of concept of responsibility I will advocate here connects seamlessly with the kind of concept of free will that I have been advocating. The libertarian version, generally speaking and as exemplified by Duus-Otterström and Smilansky, assumes the connections between these concepts are primitive moral facts. My model explains why every connection is justified and makes sense.[449] A key concept in this model is reasons-responsiveness.

[449]Cf. Dennett 2015, p. 169.

8.1 Kinds of reasons-responsiveness

Reasons-responsiveness was mentioned above (5.4.1) as a potential condition of freedom and a form of being able to do otherwise.[450] In this section, I take a closer look at the concept as it was introduced by John Martin Fischer and Mark Ravizza as a condition of responsibility. After that, I go into more detail about how it can be seen as freedom, and then how this (finally) gives a reason to connect freedom and responsibility.

Fischer and Ravizza distinguish between *regulative control*, which requires being able to do otherwise, and *guidance control*, in which one does not need to be able to do otherwise, but instead needs to be causing their own action in the right kind of way. They then argue that it is guidance control that is required for moral responsibility.[451] Based on what I have argued previously (mainly chapter 3), I will of course take the stance that being "able" to do otherwise in the sense of indeterminism is an unacceptable criterion for either free will or responsibility. However, guidance control may be relevant. It is guidance control that Fischer and Ravizza explicate by reason-responsiveness. We need to turn to the types of reasons-responsiveness they introduce next.

8.1.1 Weak, strong and moderate reasons-responsiveness

Fischer and Ravizza discuss three types of reasons-responsiveness: strong, weak, and moderate reasons-responsiveness. These kinds do not capture what I wish to say about reasons-responsiveness myself.[452] I introduce them here specifically to make clear the distinction between these kinds and what I am talking about.

Fischer and Ravizza first introduce *strong reasons-responsiveness*: If the

[450]For an overview of views that are compatibilist or close and involve something like reasons-responsiveness, see McKenna, 2011.

[451]Fischer & Ravizza 2000, pp. 31–41.

[452]Thanks to Marius Usher for highlighting the importance of this point to me.

person[453] who is strongly reasons-responsive has a sufficient reason to act in some way, then the person will recognise the reason, choose in accordance with it, and act in accordance with that choice. The important thing here is that the agent will always respond to the reason if it exists. Thus, it is no wonder that Fischer and Ravizza consider this as too strong of a requirement for responsibility; you would have to be "perfect" about it, and even if it was possible to be like that, at least nobody could ever be responsible for bad choices. This is why Fischer and Ravizza turn to a different conception of reasons-responsiveness.[454]

In *weak reasons-responsiveness*, the requirement is that, holding constant the actual kind of mechanism by which the agent operates, there is at least some possible world in which the agent recognises the reason and acts in accordance with it – or as it is put here, that there is some world in which the agent does otherwise.[455]

Though Fischer and Ravizza first examine the hypothesis that weak reasons-responsiveness is what is required for guidance control and responsibility[456], they later point to problems with it and instead introduce a new form called *moderate reasons-responsiveness*. The main problem with weak reasons-responsiveness is that of "strange patterns". In an example, suppose a person boarded a ferry, took out a sabre and killed all the other passengers. Suppose further that he would always have done that, no matter how strong the reasons, except in the case that he knew "there was a person smoking a Gambier pipe in the lower cabin." This would count as weak reasons-responsiveness, because there was some possible scenario in which this person would not kill everyone, but it would not be very meaningfully reasons-responsive, because the one scenario has almost nothing to do with any reason that

[453]They speak of a mechanism here, meaning one operating inside the person, but I leave out that level for simplicity.

[454]Fischer & Ravizza 2000, pp. 41–43.

[455]*Op. cit.*, pp. 44–46.

[456]*Op. cit.*, p. 46.

makes sense. The sabre-murderer would not be reacting to good reasons, only to a meaningless random fact.[457]

Fischer and Ravizza's response to this challenge is basically that, for real guidance-control and responsibility, the person must be able to recognise reasons in a way that is consistent and to be at least weakly reactive to them. Since weak reactiveness is enough, this means that a person may be responsible even if they are, say, weak-willed, and thus do not follow the reasons they recognise very often, as long as they could do otherwise in some cases.[458] However, to be responsible, they do need to recognise and understand reasons in a way that is regular, more specifically:

> In other words, we want to know if (when acting on the actual mechanism) he recognizes how reasons fit together, sees why one reason is stronger than another, and understands how the acceptance of one reason as sufficient implies that a stronger reason must also be sufficient.[459]

So, since I am not going to use these more specific concepts of reasons-responsiveness, I need to introduce what kind I *am* using. My concept is less precise and, partly for that reason, more relevant.

8.1.2 A different approach to reasons-responsiveness: General reasons-responsiveness

Weak, strong and moderate reasons-responsiveness are all based on having any responsiveness at all in a certain part of the system. Strong reasons-responsiveness with respect to an action applies if and only if the agent always recognises reasons

[457]Fischer & Ravizza 2000, pp. 63–67.

[458]*Op. cit.*, p. 70. See also pp. 73–76.

[459]*Op. cit.*, p. 71.

and chooses and acts in accordance with them. Weak reasons-responsiveness applies if and only if there is some possible world (any one whatsoever) in which the agent would have reacted by the same mechanism and would have done differently. Moderate reasons-responsiveness tries to address the problem of these extremes being too strong by compromising between them, but it, too, always either applies or does not, and in particular, applies the same regardless of how strong the agent's ability to react to reasons is.[460]

While it is intuitively a simple idea that concrete control requires reasons-responsiveness in the general sense of being able to act based on one's own good reasons, trying to explicate this idea with the above concepts of reasons-responsiveness gets complicated. Questions of whether you are actually likely to make the most responsive choice or not – if, for example, you only have a one in a million chance of doing it – may not be addressed at all by whether you match such a definition of reasons-responsiveness, even moderate reasons-responsiveness. A different way to put this is that an indeterministic view that is shown to be problematic for freedom by the randomness argument (see chapter 3) might still fulfil criteria of these kinds of reasons-responsiveness.

Instead of something so technical yet limited, what we need for the present purpose is to think of reasons-responsiveness in this way: the more you are such that you act based on your own good reasons, the more reasons-responsive you are in the sense that we want for concrete control. It may be necessary in some contexts to go beyond this into more detail, but this general simple idea works well enough for now. We can observe that moderate reasons-responsiveness is likely necessary (though certainly not sufficient) to have anything but very little reasons-responsiveness in this sense: someone who either does not recognise reasons according to a structured pattern *or* does not react to reasons at all is hardly overall reasons-responsive, even if they might fit some overly narrow definition (specifically, at least, weak reasons-

[460]For an overview of related problems for Fischer and Ravizza's view of reasons-responsiveness, see McKenna 2011, pp. 191–196.

responsiveness).

We can call this concept *general reasons-responsiveness* insofar as it needs its own name, though whenever I speak of just "reasons-responsiveness" after this point, you can assume I mean general reasons-responsiveness.

There remains one other question left to be answered about what kind of reasons-responsiveness my theory involves, but before answering that[461], I need to look at how and why reasons-responsiveness is very useful for explaining freedom and responsibility.

8.2 The key role of reasons-responsiveness

8.2.1 Reasons-responsiveness as freedom

Reasons-responsiveness is a promising concept for explicating freedom of will for a number of reasons that have been discussed or hinted at before. To summarise...

Paradigmatic examples of the kind of situations where (almost everyone seems to agree) free will is lacking or compromised include cases of addiction or compulsion. I explained these previously (5.4) by referring to how the agent is determined or too close to determined on a too high level: there is a general rule that they will follow that prevents them from taking into account reasons they might have to do otherwise. Thus, we can explain these kind of cases where freedom is thought to be lacking by saying reasons-responsiveness is compromised.

Reasons-responsiveness also plays a central role in the argument elaborated at length above that concrete control is threatened by indeterminism (chapter 3). Really, concrete control – *a person's control over their own choices or actions in the concrete sense that they follow reliably from the person's own motives* (as defined in 3.3.4) – is equivalent to reasons-responsiveness. The randomness argument that I presented about this is supposed to be the valid version of the general, common

[461] In 8.2.4.

argument that indeterminism threatens free will somehow. Thus, the frequently recognised intelligibility problem (see 1.1.2) about indeterminism and free will can also be said with good reason to be about indeterminism threatening reasons-responsiveness.

What about the threat from determinism? As chapters 3 and 4 establish, if we really wish to retain indeterminism as part of free will and incompatibilism, then we have to say reasons-responsiveness must be compromised to have free will, even though almost nobody wants that consequence. But if we are willing to reject this premise and try to find an explanation for why it *seems* as though determinism is a problem for free will (chapter 5), even though we do not really want indeterminism, then we have reason to look at explanations based on how determinism that is too strict and makes us not responsive to reasons is what we should really fear and maybe even what we really do fear. Thus, again: the seeming threat of determinism might just be based on the real threat to reasons-responsiveness posed by the wrong sort of determinism.

Consider also the question of how it seems we need to have different options open to deliberate between them, but how it turns out that interpreting this in an indeterministic way does not work. (3.4.9 and 5.4.4.) What turns out to be the case there is that we do not (in practice) want for all options to be open all things considered, but instead we want to be free to consider all things ourselves and make a decision based on our reasons. If the set of options were limited so that our choice functions could not choose based on it, then we would not be able to choose based on our own reasons. Yet again, what we want is to be reasons-responsive – both in the sense of the options being open for our reasons to "choose" between and in the sense that our reasons really do get to determine what we do.

This leads to the final conclusion of what the role of reasons-responsiveness corresponds to in normal view of free will. (I mean both that it is arguably what this component of typical views "really" means in some sense, and that to the extent that it is not the same thing, it is what takes the place of that component in my theory.) For both free will and responsibility, we need to be "able to do otherwise," though

the indeterministic interpretation of this concept leads to trouble. I suggest that we should understand "being able to do otherwise" as "being able to do otherwise *if there is a reason*" instead of the indeterministic version "for no reason."[462] In other words, I suggest that the idea of being able to do otherwise is to be interpreted as reasons-responsiveness. I already suggested this in the context of free will 5.4.1, and now we can see even more reasons the idea works, and it applies to responsibility as well. It has the considerable advantage that the control also required of free choices can *also* be interpreted as being about reasons-responsiveness. This means being able to understand the two things that seemed to be in opposition – the threat from determinism and the threat from indeterminism, or the corresponding hard-to-articulate positive requirements – as two aspects of the same thing.

Beyond these arguments, there is also one more reason or set of related reasons to interpret being able to do otherwise as reasons-responsiveness. This goes back to the questions of punishments and responsibility.

8.2.2 Punishments as reasons

If punishments are to act as deterrents, the case for punishing only people who are reasons-responsive can be stated very simply. In fact, it borders on the simplistic, and you would think that it takes a lot of caveats and modifications. Fortunately, it turns out that just those caveats and modifications are found in our actual intuitions and institutions.

The simple version of why it makes sense to punish only people who are

[462]I think it is fair to describe the incompatibilist, indeterminist version of being able to do otherwise like this. As argued in detail in chapter 3, in any case where indeterminism truly applies, indeterminism in the choice implies that there is always a part of the choice that happens for no reason. So, for example, if there is a reason *X* for you to choose *A* over *B*, and you do choose *A*, then it may be that *X* is the reason for your choosing in some sense – by making it more likely, say, or because your reasons only suggested *A* and *B* as possible options in such a way that your choosing only out of these options was determined by your reasons – but there is always the aspect to the situation that you might not have chosen *A* (else, it would be determined), and whether you did in fact choose *A* given that happened for no reason.

reasons-responsive is that the threat of punishment is supposed to act as a reason not to do wrong. Obviously, this is not going to work on a person to the extent that the person is not responsive to reasons including that reason. Thus, this sense of being able to do otherwise makes perfect sense as a requirement for responsibility. (Meanwhile, "being able to do otherwise" in the sense of indeterminism would not be a good criterion – for reasons elaborated in chapter 3, but shortly put because either you would have to fear you would unpredictably do a bad thing you will be held responsible, or the indeterminism would have no effect and act the same way as determinism.)

There is a further complication to the idea of punishment as deterrence that, when taken into account, only guides the system of punishment closer towards the ideas we usually have about it. If we *only* judged whether someone is to be punished based on whether that is going to have a causal deterrent effect on others, we would be considering each punishment much more "amorally" by normal standards: not whether it is fair or just, only what will happen as a result of other people knowing about it. One important part of the kind of fairness we are talking about is that people are excused when they caused something in a way that they could not help, such as ignorance (of the sort that could not be helped), disease, or "insanity" in the legal sense. However, this would not create a system of incentives of the sort that it aims to create. If a person might be punished for things they could not help or otherwise unjustly, they could not always make the choice to do the right thing and avoid punishment. Such a system would leave people having to worry about possible unavoidable and possibly unpredictable punishments, and to the extent that punishments would be unavoidable, they would not be justified by the incentive that they would give, since there would be no choice that could be taken to avoid them.[463] Two things can be noted about this line of thought: first, it hangs on reasons-

[463]H. L. A. Hart makes this same point in a context of law in 2008, pp. 40–50; I do not think there is any difference between arguments of law and ethics on just this point, and indeed the distinction is not entirely clear even with Hart.

responsiveness again, and secondly, as stated, it shows why basing the system of responsibility and punishment on deterrence and reasons-responsiveness naturally leads to the kind of ideas of justice we already have.[464]

It is true that if determinism holds, there is a strong sense in which no individual act could have happened otherwise.[465] Punishments cannot change the past anyway, only the future. If we applied the kind of model that is (roughly) the combination of libertarianism incompatibilism about individual acts and intrinsic-good retributivism, we would require for it to be possible that the agent would have done otherwise in the exact same situation in a way that would involve something like agent-causation (4.2.3), and from this, in absence of further excusing factors, it would follow that punishing the person would be right for its own sake. We have rejected this model, but it may still seem odd that it truly would have been impossible in the strictest sense for the agent to have done otherwise. Really, it is only appropriate. Though that libertarian-retributivist picture never worked in the first place, it is not really a problem that it gets denied here by its own terms too. There is no such thing as it just being right to harm someone for the sake of harming, punishment or not. The fact that the libertarian picture does not work (recall, it does not work under indeterminism either, thanks to the loss of concrete control) only serves to reinforce the point. In other words, though it would not follow that incompatibilist free will would lead to the intrinsic value of punishment, it is also illustrative to note that even if we did accept that standard, we could never get what we are supposed to get by it (cf. 6.2).

[464]Though I will not develop this point further here, this kind of reasoning can also be used to show that punishment as deterrence, when thought out properly, at least more or less supports the kinds of principles that can may be assumed to be associated with retributivism only (e.g. as discussed in Duus-Otterström 2007, pp. 129–140). Once again, the libertarian retributivist such as Duus-Otterström can only state that these requirements are primitive requirements of justice based on intuition (see 6.1 and 6.3.3 in the present work), whereas my theory can explain why they can be justified based on something else.

[465]Cf. Hart 2008, p. 42.

Recall also that there are different analyses of being *able* to do something (see 2.4.2), and indeed that taking this so far as to say it would mean indeterminism would go too far, as we would gain nothing but the loss of concrete control from it (chapter 3). That definition of being able to do different things works well to describe the kind of control over ourselves that we have reason to want, as when realised, it allows us to reliably do things for our own, good reasons. However, it also works with respect to punishment and responsibility in a slightly different sense, one that will ultimately be shown to be related (see 9.4), but is not just the same thing or even obviously very related. When we make it a rule that people in certain kind of situations will be punished if they do certain things that are wrong, it makes perfect sense to make those situations ones in which they are able to take the threat of punishment into account. Of course, this has to broadly generalise over *kinds* of situations. If we assume universal determinism and say there were no multiple possibilities in a particular situation just because it was a particular situation under determinism, no rules like this can ever apply. Besides, recursion makes it hard to say what we could even judge about such situations. (This point is related to the one in 5.2.2, though not exactly the same.) Maybe someone could not possibly do otherwise in a particular situation, but that situation was one in which the past contained the fact that we had chosen to excuse everyone because of determinism. (See 2.5.) What if, as is the aim of deterrentism, they would have done otherwise and not done the wrong thing if we had decided differently? Determinism or not, we are building the future circumstances with out present decisions.

Sam Harris states both that existing views of when people are responsible tend to coincide with when deterrence could work and that punishing people based on that would be completely against existing views of why we punish people.[466] This pair of claims can, by its nature, only be partly true at most, since it does imply we think people should be punished when deterrence does work, even if we do not put

[466]Harris 2013, pp. 60–61.

it like that. Further, we do partly put it like that when we speak of the reasons someone is or is not responsible.

I will briefly note that there exist other ideas of how punishment can have a deterrent effect. The general fact that the law is being enforced (with punishments) may serve to set an example of what is considered right and what is not, influence public values and habits, and so on, you could say generally strengthen the collective idea that some things are not to be done.[467] I will not discuss this perspective further, but insofar as such effects exist and are to be taken account in the justification of punishment, this seems to be more or less harmonious with the theory of punishment I am suggesting here. In particular, people who are not reasons-responsive may also not be responsive to effects such as these.

As before, everything that is said about punishment here, as the most extreme example, applies equally well to other practices of holding responsible, such as praise, blame and reward. All of these can have a justification based on their effects of behaviour: even if you will not be outright punished, you could react to the knowledge that one choice will be praised and the other scolded. This is, again, reacting to a reason, and thus is can only be expected of those who are reasons-responsive.[468]

[467]Yli-Hemminki et al. 2022, pp. 28–30.

[468]A potential mismatch between the deterrentist rationale and common intuitions and practices may be when successful attempts at crime are punished more severely than unsuccessful ones, and negligence causing an accident is similarly treated as more severe than negligence merely risking it. H. L. A. Hart argues this may be due to our tendencies of retributive thinking, but not in accordance with retributivism any more than it is with deterrentism. (2008, pp. 129–135.) As far as deterrence goes, there is, of course, reason to punish attempts in order to deter wrongdoing, and similarly for negligence. (*Op. cit.*, pp. 128–129.) I do not know that this is something that needs to be changed about the justice system, though, unless it were specifically shown that doing so would achieve a great deal more deterrent effect. Since punishment is a necessary evil rather than an intrinsic good, then if we can get away with inflicting less of it in a way that has internal consistency up to a certain point, and that seems acceptable to people, without sacrificing much deterrence, we may be achieving a better outcome by acting like this than by being more fully consistent while paying the price of more punishment. If anything, if it were possible, we should look for more opportunities to slack on punishment if we can do so without contradicting

8.2.3 Punishment vs. treatment and the (non-)essence of disease

What do we do, then, with people who are not responsive to the threat of punishment and other aspects of being held responsible? In some cases, such as children, we let them be under the control and supervision of someone more responsible. In other cases, we consider them something like sick and in need of treatment. Though this is another thing that just feels natural, the point of responsibility and reasons-responsiveness explains it, too, perfectly. If we cannot hold someone responsible because they are not responsive to that, we can instead try to treat them so that they become responsive and able to function in society. We excuse them from judgement and punishment – which are only justified as a necessary evil anyway, and in these cases cannot be so justified – and instead try a different method they might respond to.[469]

What does it mean for a condition to be counted as a disease? What is the difference between treating, say, alcoholism as a sickness (mental illness) or treating it as a vice? Diseases are of various kinds, and as far as I can see, the common factor between them is functional. When somebody is ill, their condition is such that it makes sense to excuse them from some of their normal responsibilities.[470] This might mean letting them skip coming to work or to treat their alcoholism as something needing treatment rather than blame. In any case, this logic works just the same way as my proposed explanation for why when treatment is seen as the appropriate response instead of blame.[471] Naming something as a disease can also be

common feelings of justice and without compromising deterrence.

[469]Cf. Strawson, n.d.

[470]Of course, some things are counted as mild cases of disease that do not really excuse one from anything, because they are milder versions of the same kinds of things that are counted as diseases on such a basis, such as a very mild cold or migraine.

[471]Cf. Duus-Otterström 2007, p. 179.
Cf. Baumeister 1996, pp. 297–298 for a more negative side of how just about anything can be defined as mental illness.

empowering via separating it from the person's "self"[472]. It will be argued later (8.3) that we cannot simply make a clear distinction between what is part of the self and what is not, and that is instead a stand-in for something that needs to be analysed differently, but such a way of drawing the distinction does make sense in specific contexts in life as opposed to metaphysical analysis.

8.2.4 Sufficient reasons-responsiveness

We have now established that the theory developed in this work has general reasons-responsiveness as a central component of free will and of what is required for responsibility, and that there is reason to require for people to be reasons-responsive before holding them responsible. This, however, raises a further question of just *how* reasons-responsive they need to be. General reasons-responsiveness was introduced as a compromise concept that is not too strict or loose, but it is inherently a scale. So, just how reasons-responsive on that scale should a person be to be held responsible?

If we required an extremely high level of reasons-responsiveness, something like strong reasons-responsiveness, nobody who did something wrong would be responsible for it insofar as there are always reasons not to do things that are wrong. (This may seem like a leap, because after all people are responsive to their *own* reasons, but I will later argue that morality, reasons and freedom are interrelated: see 8.3.2.) Besides, nobody actually is that reasons-responsive; that would make them a perfect being like Reid's God (see 11.9).

There is not some metaphysically or logically mandated answer as to where we draw the line. All we can say is that people should be held responsible insofar as

---

The point has also been made that we pathologise things insofar as they are treatable and when they become treatable (Solomon 2002, p. 33). Though this is not the same as saying that that which *cannot* be handled *without* treatment is pathologised, it is close and probably harmonious with the same kind of reasoning I present here.

[472]Solomon 2002, pp. 463–464.

they have a suitable normal adult human level of general reasons-responsiveness that makes them susceptible to such feedback.[473] I will call this *sufficient reasons-responsiveness*, though, as with general reasons-responsiveness, you can assume this is what I am speaking of if I just speak of "reasons-responsiveness" in the context of moral responsibility (and free will) after this.

Note that sufficient reasons-responsiveness as a criterion is supposed to be used in making generalisations, as in law. There are various situations where we have to make judgement calls about the degree of someone's responsibility for a particular thing in a unique situation (see especially 9.5), and even courts interpreting the law on particular occasions do this.

8.3 Searching for the "real self"

There is one more problem left. What is the difference between a compulsive, unfree desire and one that is merely strong and constant? This section gives an answer to that, starting from not supposing any motives to be more valuable than other to start with and then showing that we can still make distinctions between them by appealing to a necessary requirement of harmony. These points can be applied equally well to free will and to responsibility, and here for once I will stop making the distinction between having free will and being responsible.

8.3.1 The importance of the "self" for freedom: Responsive to what reasons?

I argued previously (5.4) that a thing like alcoholism or kleptomania is counter to freedom because it causes loss of concrete control because it makes one unresponsive to reasons a person has to do otherwise. This is actually not enough.

[473]Cf. Dennett 2015, p. 104, Reid 2010, p. 233. See also, for illuminating discussions of beings such as non-human animals and children that may be lower on a spectrum of the same kind of moral agency or personhood, Schlossberger 1992, pp. 73–77 and Wise 2005, pp. 32–33.

The point was that the rule in the lines that "She will always shoplift" does not need to be literally deterministic, only close to it. If that is the case, how is it different from a merely strong conviction? Let us suppose that Martin Luther from the previous examples (see 2.4.2 and 5.4.4) is free and responsible when he says he can do no other, because he is following his own strong principles. Perhaps, then, these principles would also cause him to act predictably in certain ways in many situations, unless there was a really strong reason to do otherwise. Why is this different from the case of addiction?

The difference comes down to this: we do not count addiction as an interest, but we count a strongly held value as one. We could also see some values as more worthy than others, only compounding the question. To make judgements like this, we need to be able to make some kind of a distinction between what motives are really a "part of the self," or otherwise more worthy, and what kind of ones are not. It may be intuitively easy to say that an addiction is not part of the self, whereas something recognised more traditionally a value is, but we need something more than that.

One relatively obvious direction to look for answers in this respect is to make a distinction between those "inner" motives and influences that are really part of the person's self and those that are not. Susan Wolf has called these *real-self views*[474]. Thus, perhaps an addiction, especially one that the person does not embrace, is not part of the person's real self. In such a case, the person could be more free if they could be controlled more by their real self and less by the addiction. Certainly (even with hindsight after all the problems I am about to raise), that is true in *some* sense. But what sense, exactly? How are we going to distinguish what is part of the real self and what is not – or is that even the right question to ask?

One attempted answer is the appeal to second-order desires, which are desires

[474]Wolf 2002. In this text, she also introduces an answer to their problems that shares some features with the one that I introduce in the next section, though she still has to appeal to a concept of what is right, so it ultimately faces the same problems of circularity that I discuss next: how are we to decide *what* is sane or right?

about which desires to have. Harry G. Frankfurt famously described this view in his article "Freedom of the Will and the Concept of a Person". I will examine this view and the problems facing it as an example. In Frankfurt's view in this article, a person has free will insofar as they act according to not only their desires, but those desires that they *desire to* act according to. Thus, free will involves *second-order desires*, which are desires about which *first-order desires* to have. First-order desires are ordinary desires to do things. A further distinction to be noted is that those first-order desires that the person is actually moved by are *volitions*, and second-order desires about which first-order desires the agent wants to be moved by are *second-order volitions*. If someone were to have second-order desires about having some first-order desires, but not a desire to be moved by those desires, that would be besides the point for their freedom. So: a person has free will insofar as they act according to their second-order volitions. Someone who is driven by addiction would be unfree insofar as they did not wish to be driven by it, though someone consciously embracing the addiction might be free.[475]

Several kinds of difficulties face this account, many having to do with the problem I am discussing here of needing to know which motives to consider as part of the self; just anything targeted by second-order volitions will not do. Again, I only look at some examples by the way of leading up to my point, based on a critique by Dennis Loughrey.[476] Loughrey raises the dilemma that it seems that second-order desires both have to be instrumental and cannot be that. If they are formed on the basis of rational reflection on first-order desires, it seems to be implied that they are instrumental for fulfilling those desires. You would form a desire to, say, not be addicted to a drug because on reflection, would see that your first-order desire to remain healthy requires it. However, if second-order desires merely reflected first-order desires, they would not supply anything that would make them truly more part

[475]Frankfurt 1971.

[476]Loughrey 1998.

of the person than the first-order desires themselves.[477] However, if second-order desires were *not* instrumental like this, they would just be desires we would happen to have, and again, there would be nothing special about them making them more truly our own than other desires.[478]

Loughrey ends with a solution of his own that does not answer my question any better. He writes that it would work to say that what makes the difference is rational approval. If we say this in terms of second-order desires, then second-order desires would be an indication of the authentic self due to being ones that we rationally approve of. However, he contends that in this case, it is redundant to speak of *desires*; why not just speak of rational approval?[479] Now, from the point of view of the question as I posed it, it does not matter whether we speak of second-order desires or rational approval. The same challenge still arises: what makes rationally approved desires so authentic? Further, *rationality* already involves a value judgement. How do we judge what is really rational? It may be a very good heuristic for finding good, authentic solutions to deliberate on them carefully, but if it is taken as definitional instead, one could in principle end up being authentic and free with the most insane principles[480].

My answer to this problem in the next section is based on seeing how instrumentality itself can lead to something relatively objective by which authenticity can be judged. There is no fundamentally real self, nor things clearly not belonging to the self, but given everything that could be in the self, less vaguely every desire we have, already gives a basis for normative judgements about what parts of oneself one *should* endorse, and how doing so makes one more free.

---

[477] *Op. cit.*, pp. 216–221. I have simplified the argument somewhat.

[478] *Op. cit.*, pp. 221–222.

[479] *Op. cit.*, pp. 222–225.

[480] Cf. Wolf 2002 again.

8.3.2 On the nature of morality – and free will

> Every natural appetite, desire and affection, has its own present gratification only in view. A man, therefore, who has no other leader than these, would be like a ship in the ocean without hands, which cannot be said to be destined to any port. He would have no character at all, but be benevolent or spiteful, pleasant or morose, honest or dishonest, as the present wind of passion or tide of humour moved him.
>
> –Thomas Reid[481]

As promised, I now propose a solution to the question of how we can determine which motives we should consider our free interests that does not assume some interests are real and some are not, nor that some are better than others, but instead starts by treating them all equally. This account overlaps with a theory about the nature and justification of morality in general that I have sketched out in an essay previously[482], but I will not dig too deeply into all that. Though I combine it with other ideas in the essay, the part of the answer I present here is from Mary Midgley, who also attributes the basic idea to Charles Darwin. As far as explaining the nature of morality is concerned, one part of the question has to do with explaining why moral truths have the peculiar features they have. Some among these features include that judgements about one should do morally speaking (all things considered) are (arguably) supposed to override everything else, and that while they are so supposedly final and somehow action-guiding, it is still possible for one to make a moral judgement about what one should do and still not do so.

For Darwin, there was a related question about why humans are moral beings while other animals are not. He takes the example of an ape killing another one in a fit of rage and then acting vaguely regretful about it but lacking the ability to realise

[481]Reid 2010, p. 150.

[482]Kokko 2018.

that taking such an action in the first place would be a bad idea regardless of how strong the rage-filled impulse is.[483] By contrast, he thinks, a human being has reached a level of intelligence where they can think in terms of the future and past and consequences of their actions – and thus a human being has both the need and capacity to do something more. We cannot simply go by the strongest inclination at each moment, since it may contradict our other interests, even if they are not so strongly felt at that moment. This is especially true considering that we often have strong but short-term impulses and weaker but long-lasting inclinations – something like that ape who feels like killing someone at one moment but would overall prefer to have that other ape around. Given that we are like this, Darwin and especially Midgley conclude, it is necessary for us to make judgements about what we should do in a given situation all things considered – and, likely, to have constant rules by which to make these decisions.[484]

If we need to have rules and judgements like this, those rules and judgements already have either all or at least many of the properties of moral rules and judgements. Consider the point above about moral judgements being supposedly overriding but also not always followed. It is the whole point of judgements such as I just described to be overriding, so that, e.g., even if you are strongly feeling like attacking someone at the moment, you still know it is wrong. At the same time, though we have a practice and system of making such judgements, there is no guarantee that we act in accordance with it. We can make those judgements by that system or those rules but still be more influenced by something else, probably the

[483]We will put aside the question of how accurate this is as a description of ape behaviour. For some reason, people like to make generalising statements about non-human animals and what they lack when they are really talking about humans and using the others as foils. Darwin is at least using it as a conscious strategy and presumably trying to base it on real ethological observations, but regardless of how well he does that, the current discussion *is* about what humans are like.

[484]Midgley 1994, chapters 13–17; Darwin 2007, chapter IV.

momentary impulses.[485]

An interesting point about this view of the necessity of morality, one that will also be relevant when we get back to applying the same ideas to the self and free will, is that it is an individualistic version of a simple and powerful argument for the necessity of morality in societal and interpersonal relations. Shortly put, when different people are interacting with each other and making use of the same pool of resources, if everyone tries to fulfil their competing interests without regard for the others, everyone ends up losing, because they have to spend resources on competing for resources. It makes sense instead for everyone to agree to rules that apply equally to everyone, whereby everyone can have no more than their own share, but they still get more than if they wasted effort and took risks to compete over everything.[486] (This is a version of the same "prisoner's dilemma" reasoning mentioned in a different context in 7.4.2.) Hence, the necessity of overriding moral rules to harmonise everyone's interests and keep them from getting in each others' way. Now, the idea by Darwin and Midgley is an intra-individual version of this same idea in that it harmonises the competing motives within the individual in the same way. Instead of pursuing one goal at one time and another at another time, hindering both goals, we can follow these overriding rules and judgements in pursuit of a harmonious set of goals – at least in principle, and even doing so imperfectly will

[485]Alan Donagan (1987, chapter 8), makes a similar point about it not being enough for agents to only have desires of various strengths, but he ties it to "will" as the factor as it were choosing which inclination to follow. This might run into trouble with asking for which inclinations and choices are to be identified with the will, however, the same as with what is part of the real self. It also gets close to the discussion about willpower, which is a whole can of worms of its own (see e.g. Koi 2024, Sapolsky 2023, chapter 4).

[486]Midgley mostly does not draw this comparison but instead pits such a contractarian explanation about the origin of morality against hers (in another source). This is because that theory, as she treats it, starts from the assumption that human beings are selfish. (Midgley 1993, pp. 30–36.) As I make the point here, it is not dependent on humans being selfish about fulfilling their needs but simply about having conflicting needs. Midgley does come close to acknowledging this parallel on page 41 in the same article.

almost automatically be better than having no values.[487]

As a further point, this idea also provides criteria for choosing between competing moral rules. Some moral rules are instrumental, and thus it is open in principle to question whether they really help reach their goals or not. Still, some rules have to be taken as fundamental, as that towards which all instrumental values ultimately try to lead. While there can be no logically valid way of choosing the right values, as long as we admit something like happiness and the absence of pain as the most basic value, we can judge all other values based on whether they are really in harmony with that value. We can also see if some ultimate values contradict with each other, causing inevitable disharmony and strife, and judge them on that basis.[488] This, in a way, answers the dilemma posed by Loughrey for Frankfurt, too: moral values (as second-order desires of a sort) are instrumental, but in such a way that they need to be elevated to a position where they are considered as far more important than that, overriding even individual intrinsic values in the service of all intrinsic values collectively.

The point I am getting at here is that we can do this same thing when we consider which interests to count as appropriate for free willing. An unwillingly addicted person, for example (insofar as the addiction is so wrong they are really considered not free), is presumably driven by an uncontrollable impulse at one moment but regrets it later. They are not responsive to the kind of reasons that are themselves responsive to the person's reasons in general, such as the judgement that drinking so much is overall bad. Meanwhile, even a willingly addicted person might be deluded about what interests or potential interests they have that are going

[487]David Wong states the same point very directly: "Such an argument could be supplemented by an explanation of why human beings have such a thing as a morality. Morality serves two universal human needs. It regulates conflicts of interest between people, and it regulates conflicts of interest within the individual born of different desires and drives that cannot all be satisfied at the same time." (Wong 1993, p. 629.)

[488]Midgley 1993, pp. 41–42.

unfulfilled in their hyperfocus on their addictive interests. Of course, we cannot simply dismiss the possibility that there *are* willingly addicted people who genuinely have things well in their lives, in which they might be fairly free, though perhaps potentially lacking reasons-responsiveness in situations where the need to do otherwise would arise. Meanwhile, someone like Luther unwaveringly following their moral convictions might be doing the right thing all things considered, though that assumes their moral values are ones that would be judged right by this logic of morality as involving harmony. A closet homosexual fervently crusading against homosexuality for no other reason than cultural hang-ups might be meaningfully said to be made unfree by their having such a moral value (which those making such a judgement would then not recognise as a correct moral value). They might still be different from someone with addiction insofar as the addicted person's ability to modify their behaviour was more limited – after all, the idea of someone following their morality as rigidly as an addicted person follows their addiction is more of a possibility than a typical case – but it is possible in principle that an unhealthy moral idea followed very rigidly would be just as limiting and unhealthy as any addiction.

Thus, we have asked the necessary question of which reasons should count as the agent's own good reasons that should be followed in free will, and which kind of ones should not, and we have come to the conclusion that we can judge between these by looking at what promotes overall harmony in the system of the person's interests.[489] What promotes such harmony is also what best promotes those interests overall. Due to the overlap with the justification of morality, this also implies that the hypothetical perfectly free agent would also be perfectly moral (according to the metaethical theory I have sketched).[490] This is part of the reason why I earlier defined sufficient reasons-responsiveness as the base level required for basic moral

[489]This idea also resembles Eugene Schlossberger's idea of moral personhood as having a coherent evaluative framework about the world that is expressed in one's interactions with the world (Schlossberger 1992, pp. 33–34).

[490]Cf. Schlossberger 1992, p. 37. See also 11.9 in the present work.

responsibility (8.2.4) – so that people can still be held responsible for wrongdoing even given this. The perfectly free agent is in any case something that we can only approach to being, never reach.

8.4 Conclusion to chapter 8

The previous chapter argued that we cannot justify differential moral treatment of people based on a concept of and ideas about responsibility that are based on only intuitions, no matter how strong. This chapter has has argued that we do not need to base our notions of responsibility on intuitions, and a notion of freedom and responsibility based on reasons-responsiveness and considering the consequences of choosing these values reaches the same basic conclusions – often remarkably closely just the same ones.[491]

In his famous article "Freedom and Resentment"[492], P. F. Strawson argues against both incompatibilism and a utilitarian justification of our practices of holding responsible: Determinism does not threaten our practices holding people responsible, but those practices are not saved by their mere utilitarian usefulness. Instead, they exist because they are part of our social psychology. When we regard people as proper moral subjects,[493] we also hold certain reactive attitudes towards them and their actions.

Just to state or argue (or even prove, if that had been done) that this is how we work is not itself an ethical justification. What if it was simply *wrong* to act in accordance with these attitudes, or some of them? What if that was the case, but they were still humanly unavoidable or very hard to avoid? This would raise many difficult lines of questions. Fortunately, we do not need to go down these paths at

[491]This closely matches what Manuel Vargas describes as denotational revisionism and the reasons it can be justified (Vargas 2011, p. 462. See also p. 467.)

[492]Strawson n.d.

[493]See also 5.3.4 in the present work.

this point. I have already argued that our intuitions about responsibility and justice correspond to a system that we can morally enforce for other reasons. To whatever extent Strawson's vaguely empirical thesis is true, it is itself another practical justification for maintaining our current practices: they are also very natural to us, meaning that people are prone to thinking and acting in accordance with them.

As we have seen previously, for those putting much value on intrinsic values (and intuitions), consequentialist arguments for moral values tied to responsibility seem like a cop-out. I have argued in this chapter that these values are bound to human social and societal existence in a number of ways, of which the consequentialist and deterrentist ones are an important kind, but which also tie to our needs, practices and emotions in other ways. I at least partly agree with Strawson that something is missing from the consequentialist picture, and that something is provided by understanding how holding responsible is and needs to be a part of our practices and attitudes towards each other.

Even when philosophers defend something as an intrinsic value, they tend to acknowledge its ties with other morally relevant values. Still, insofar as something is fundamental, it is standing on nothing but itself, and if it is considered fundamental because of intuition, it is really standing on intuitions. I find it much more convincing to base moral values on an understanding of why they are needed. This means they are based on the very structures of our lives. Even if we do not feel like we should accept some given principles, we may suffer if we do not follow them in our lives if those principles are ones that follow from what our existence is like. That is what it means for values to be significant.

I started with free will, analysed the discussion about it, considered the options, and sketched my own solution for how to understand it. Next, I did the same with responsibility, ending up with something compatible with my view of free will. All I have left to do now is to bring these pieces together more completely – and then say something about what this means for us as human beings faced with such concepts of free will and responsibility. That will be the topic of the last chapter.

9 A Unified Theory of Freedom and Responsibility

> Developing a comprehensive theory of free will goes beyond the scope of any single piece of work and can only be achieved by a community of scholars, across multiple disciplines.
> –Christian List[494]

With reference to the above quote, my theory may not be that comprehensive, and certainly a lot of work by others has also fed into it. However, I will say this all fits together surprisingly well. In this chapter, I will show how, after supplying the last few missing pieces.

9.1 How to ask and answer ultimate questions, part 2

In section 1.4 near the beginning of this work, I talked about meta-level questions related to how one should go about asking the kinds of philosophical questions that relate deeply to our lives. Here, I look at such questions again, this time with the background of the discussion that we have already had in this work.

9.1.1 Not metaphysics after all

The questions of determinism, free will and responsibility are often posed as a metaphysical question of freedom. This question depends on a metaphysical question

[494]List 2019b, p. 10.

about determinism, and answering it leads to a metaphysical answer about responsibility. Then, ethical answers are based on these metaphysical ideas, based firmly on intuitions that they are.

As I have progressed with my analysis of the question here, paying attention to what is important, this study has quite intentionally moved further and further away from such an approach. The discussion early on has been metaphysical, but the conclusions reached by that discussion has led away from considering metaphysics as key. The questions of free will, agency, responsibility, justice and so on are really questions about our lives and how we should live them. Metaphysics is much more limited than it is often thought to be in its ability to contribute to such a discussion, let alone to define it. Though this might be contrary to common intuitions in some sense, things do not become more important if we suppose that they just are important than if we find reasons in our concrete lives for why they are.[495]

9.1.2 Solipsistic choosers or a part of nature?

> If we are responsible, and if what I have been trying to say is true, then we have a prerogative which some would attribute only to God: each of us, when we act, is a prime mover unmoved.
> –Roderick Chisholm[496]

Alan Watts argues in his book (called *The Book: On the Taboo Against Knowing Who You Are*)[497] that the western focus on the individual sells us an illusion that each of us is separate in three ways, each of which is impossible: separate from other people and society, separate from physical nature, and metaphysically separate from the rest

[495]Cf. Dennett 2015, pp. 172, 178; Harris 2013, p. 57.

[496]Chisholm 2002, p. 55.

[497]Watts 1966.

of the universe. He does not mention free will as one such area[498], but incompatibilism could be added as a fourth one.

Consider what we are left with if we insist on an incompatibilist perspective of what free will means if the arguments for compatibilism in this work are right.[499] Determinism does not stop us from following our own interests, deliberating, resisting temptations, being creative, or holding ourselves responsible in a way that functionally works. It is indeterminism that threatens to cut the tie between our choices and our reasons. Yet, to be free, according to the incompatibilist conception, our relevant[500] choices have to be separated from their antecedent conditions – not only the world that we have lived in all our lives, and the families that we come from, the genes that are part of us, the environment we are embedded in, but even our current desires and reasons. All of these can have some effect, but ultimately for our choice to be real, it has to be severed from all of these. We must make the choice as solipsistic individuals, so independent that we are parted from everything that makes us up, both outside relationships that are vital to us and things that are the most intimate inner parts of ourselves. Everything must be stripped of the power to determine us, even through an internal process of all-things-considered deliberation, or (if that sounds too rationalistic) the strongest feelings and values we have.

This conception of freedom is an extreme version of negative freedom, freedom from constraints. It is this at the expense of positive freedom, which means having the opportunity and resources to actually realise your goals. The consequence argument embodies this with its assumption that things are not up to us if they have prior necessitating conditions. This might seem natural enough, until we realise that something like prior necessitating conditions partly constitute what it means for us to act at all. By denying them, we are denying the world and our selves. Calling this

[498]Though see Watts 1966, p. 89.

[499]See especially 4.3.2.

[500]See 3.3.6.

freedom is like, if we had nothing to eat, to call it "freedom from food".

To show an alternative to this, I will try to do something different than the argumentation I have given so far: I will paint a picture of the alternative way of looking at the world.

We are born of this world and are part of it. Though there are ways in which it makes sense to look at ourselves as being individual, unique foci of action, we can hardly deny that we are part of the world. We have come to be here because of our parents, because of evolution of life, because of how the solar system has formed the way it did. We are able to go on living and existing thanks to a constant exchange of matter and energy with our environments. Our needs, desires and interests emerge as a combination of how we are built inside and that interaction with the environment. We are in the universe and the universe is us. It is so much more than just conditions that prevent our existence.

I know that in some sense, every achievement or quality of mine that I am pleased with and that is perhaps part of what I consider important for my identity, all of these are in some sense born from factors outside myself. When I think about this nowadays, it does not make feel like they are taken away from me. It makes me feel like a part of the universe, an expression of the emergent pattern of life and purpose struggling and flourishing within it. All those flows of causality and structure from ancient times coalesce in me, among innumerable other individuals. All those flows of human culture, too. What I am as myself is a part of a greater whole – I need not be an impossibly solipsistic, self-creating, isolated individual to own my own acts, qualities, and achievements. I am essentially both an individual and a part of something greater, and being both is just what I wish for.[501]

[501]For an articulation of the opposite point of view and how Smilansky's ultimate perspective leads to such nihilism, see Smilansky 2011, p. 435. Though nothing keeps Smilansky from claiming my point if view is illusory (though my arguments in 6.3.3 work indirectly as arguments against such a claim), I for my own part claim my perspective is an example of how Smilansky's ultimate perspective is not uniquely the deepest perspective.

Now, with all this said, the perspective of the agent making choices as a locus of action rather than a mere node in a web of causation does not need to be dismissed as a mere misguided illusion. Just as I am explaining free will as something that does not need libertarianism but can have the same kinds of features without indeterminism, and responsibility as something that is not a moral primitive or justifiable by intuition but ends up working much the same way when justified, I can see a point to the agent-centred perspective. The primacy of the agent as the origin of the choice makes no sense as a fundamental metaphysical law that overrides natural causality, but it does make sense as a perspective that tells us something true about how the world works while appropriately leaving out some other aspects. It is valid to conceptualise parts of the world through the perspective of agency (5.3.3). Even if choices are deterministic, it is true that it does not make sense to conceptualise them as predetermined (2.5), that it is natural enough to feel that different options are open for us (5.2.1), that we could not make choices if the options were not open in *some* sense (3.4.9 and 5.4.4), and that we are meaningfully responsible for our actions. The chain of causality and possible determinism involved in our choices ultimately only happens by passing through *us*. It does not pass us by; each normally taken action is an event where an agent plays a crucial final part. Each agent is a part of the universe, yes, or else we run into absurdity and possible alienation, but each agent is also a meaningful nexus in the chain of causality. A perspective highlighting the latter aspect tells us something true and important about how the universe works, as long as we do not mistranslate that perspective into a metaphysics that denies other equally important perspectives.

9.1.3 Why there is no need to ask whether determinism is true

I return now to a different side of the question introduced near the beginning of the dissertation (2.1.5). To recap, it might have seemed important, in a study of "determinism, free will and responsibility" to know whether the universe is in fact deterministic, or even to what extent. However, I ended up arguing that the basic

idea of incompatibilism that determinism in general precludes free will is one that should not be accepted (chapter 4). As such, the most obvious reason to need to know whether the universe is deterministic at the bottom ended up being unnecessary. We could not conclude based on that knowledge that that alone precludes free will.

Instead, I concluded that a kind of determinism is required for free will (chapter 5). However, this kind of determinism is a relatively specific sort that appears on a high (but not too high) level of description, and it has also been established that determinism on one level of description can coexist with indeterminism on another (Appendix A). Thus, finding out that the universe is indeterministic at the bottom would also not preclude or even threaten free will in the sense that I have argued to be relevant.

In addition to this, the kind of free will that could not exist with any amount of indeterminism on the relevant level is merely the ideal form of free will. When I have been making the argument that indeterminism always contradicts concrete control to the extent that it has any effect on it, the point was mainly to show that the idea that indeterminism is needed in principle is seriously flawed (chapter 3). If indeterminism in the context of free will can, when properly analysed, only either give nothing or give something negative, then we are left with nothing but an intuitive sense of needing indeterminism that contradicts what else we want (chapter 4). That is the point I have been trying to make. The point is *not* the mirror image of the incompatibilist point that if we find the world is deterministic, we would have to conclude there can be no free will. Thus, I have not been trying to say that the world must be deterministic on the right level or we have no free will. We are unlikely to have perfect free will in any case – in fact, it strongly seems empirically impossible (see 9.6.1) – but free will is a matter of degrees anyway.

Thus, the questions we should be looking at within the reality we know are not about whether the universe is deterministic or indeterministic, not really even on a higher level where our decisions happen. Instead, we need to look at the specifics on such relevant higher levels. How close are we to being responsive to reasons and rational in the right way? I make a very quick start at answering such questions in

9.6. It is questions like this that we should really be looking at to find out our degree of free will, after, of course, first seeing what free will even means and which meanings we should examine and adopt. It is premature to ask about determinism before that, and, it turns out, it is unnecessary to ask about it after.

9.2 The definition of free will

Finally, after all this work to find out how we should define "free will", I give you the definition that I propose for use, as part of the theory that I am advancing. I will be brief here, since though this definition needs much explanation to be really understood, that is something I have been doing throughout this work. I will, however, discuss the implications of this definition a little after giving it.

The definition of free will we have gradually arrived at is all about being able to truly act from one's own best reasons. It is also a spectrum, not a binary choice; one can have more or less free will, and to have complete uncompromised free will is a hypothetical godlike ability. The definition could also be applied either in particular situations (or kind of situations, or set of situations) and to a person in general, although the latter is bound to be vague. With all that said, here is the definition:

> An agent *A* has free will to the extent that, when making choices, *A* decides in accordance with *A's* interests all things considered.

The connection with reasons-responsiveness is obvious, and hence, if you want to review the arguments for why this definition matches what has been observed about what is required of free will, the best summary is in the section "Reasons-responsiveness as freedom" (8.2.1). What is added to that here is "all things considered," referring to the discussion just before this about how we need to identify the "right" interests and how that can be done by referring to the harmony of the whole (8.3).

As one more note on this formulation, it is overall perfectly natural and

sensible to speak of *A*'s "ability" to make the best choices, but I do not say that in this precise formulation because that would require a further potentially controversial analysis of "ability".

9.2.1 Freedom and rationality

Free will as presented here has a lot to do with rationality – even to the point, perhaps, of just being a variety of it. The central thing about free will in my definition has been general reasons-responsiveness, and that means responsiveness to all relevant reasons. This is basically equivalent to instrumental rationality, and when it is expanded to include considerations that justify which values should be selected, it also encompasses rationality in a normative sense.

The connotations of these words may sound repulsive from a certain point of view: What, instead of freedom, I am only offering rationality? Well, this is where we end up. Concrete control and thereby reasons-responsiveness has been argued to be a good measure of both the control and the alternative options required for free will. Instead of the connotations, consider the analyses. Instead of the words in isolation, consider what has been said as a whole. Remember also that the topics related to creativity have not been discussed properly, merely to show that they are not in contradiction with determinism (2.6.3 and 3.5). There may be a lot to say there, about freedom as spontaneity perhaps, that would sound much less rationalistic. Also keep in mind that the rationality discussed here is not rationalistic in the sense of adhering to a mere *stereotype* of rationality as cold reasoning.

9.3 The definition of (moral) responsibility

Given what has been discussed so far, I can give both a definition of moral responsibility within my theory and a simple justification for why it is morally right to apply such a concept of responsibility and act accordingly.

An agent *A* is morally responsible for morally relevant state of affairs *S* (past,

present, or future) to the extent that *A* had/has/will have a reasonable chance of affecting whether *S* occurs or not. *A* is morally responsible for an act *B* to the extent that *A* has a reasonable opportunity to do or not do *B* and that *B* affects whether *S* occurs.

This applies to role responsibility, outcome responsibility and liability responsibility (see 7.1).

This definition of responsibility ethically justifies itself via the following simple reasoning: It is right to ask or demand an agent to act in accordance with what will bring about the morally better consequences, to the extent that the difference in the consequences is significant and, importantly, to the extent that doing so does not ask too much of the agent for other reasons. There is no need to combine a metaphysics of agency with ethics via a link supposed to be a brute fact.

The above reasoning seems to make sense for all kinds of role responsibility, not just moral, and probably other kinds of outcome responsibility and liability responsibility as well. If you want *X* to get brought about, it only makes sense to hold responsible for *X* a person who is reasonably able to bring about *X*. This same logic both ethically and otherwise applies in the question as it is discussed in section 9.5 below.

Next, we will see how this concept of responsibility fits perfectly into the scheme I have presented of how freedom, responsibility and other concepts are intuitively related.

9.4 Filling in the schemata of freedom and responsibility

Now, we can look at the schematic features of free will and moral responsibility introduced in 7.2 and see how they are justified in my theory. Recall that we will be using the following definitions and ideas:

- "The ability to do otherwise" is interpreted as the aspect of general reasons-responsiveness that you can (will) do otherwise if there is a good reason to.

- Free will is as follows: "An agent *A* has free will to the extent that, when making choices, *A* decides in accordance with *A*'s interests all things considered."
- Moral responsibility for *X* is explained as it being right to morally demand *X* of the agent in the situation, which is only true insofar as it is reasonable to ask for the agent to do it.
- The ultimate purpose and justification of praise, blame, reward and punishment is to encourage good behaviour and to prevent harmful behaviour.

1. **Having free will requires the ability to do otherwise.** This is true because deciding in one's best interests all things considered requires the ability to do otherwise if there is a reason.
2. **Being responsible requires free will.** This is true because if one is not capable of deciding according to good reasons, it is not reasonable to say they are morally obliged to do something because it is good.
3. **Being responsible requires the ability to do otherwise.** This is true because if one is not capable of avoiding making some decision even if they have reason to, it is not reasonable to say they are morally obliged to make some other decision.
4. **Being responsible for a good or bad act implies being liable to good or bad treatment: reward, praise, punishment, censure.** This is true because if it is reasonable to ask something of someone, then insofar as it is needful enough to enforce a kind of behaviour or its absence, people who might choose to act one way or the other should know that there are consequences to it and be encouraged or deterred.
5. **Being responsible implies the absence of excusing conditions, which include at least: ignorance (within limits), lack of understanding, coercion, absolutely or relative inability to do the right thing, belonging to an excused group or category (non-humans, children, the "insane"),**

and being determined in an unusual way such as an illness. This is true because responsibility requires the reasonable ability to make a difference to the thing you are responsible for, and all these conditions compromise or exclude it in various ways. Ignorance and lack of understanding: you cannot be reasonably asked to avoid what you do not even know follows from what you do. Coercion: besides it not being reasonable to ask people not to be coerced, it is much better to hold the coercer (role) responsible, as otherwise coercion gives a way for people to get something done without being held responsible for it. Absolute or relative inability to do the thing: this directly means that it is not reasonable to ask the person to do the thing. Belonging to a group such as being "insane", a child, or non-human animal: these groups are not generally responsive to reasons at the same level as normal adults. Being determined in an unusual way: Again, it is not reasonable to ask someone to do otherwise if it is about something that their body or mind causes outside the control of their deliberative and potentially reasons-responsive mechanism.

Similarly, and partly overlapping with the above, we can explain the connections between notions of responsibility explained by Vincent[502] that were discussed in 7.1 based on this theory, mainly on the simple idea that the agent must have a reasonable chance of affecting whether the state of affairs they are held responsible for. Outcome responsibility requires causal responsibility because one cannot affect something one plays no causal role in, and role responsibility because role responsibility is practically defined based what it is reasonable to ask for. Role responsibility requires capacity responsibility because one cannot reasonable affect what they have no capacity to affect. Liability responsibility requires outcome responsibility because the differential treatment implied by holding liable is only

[502]Vincent 2011.

justified by a system of feedback that requires reasons-responsiveness to work.

Of course, none of this is completely novel or never having been said by anyone else. I do not mean to boast too much when referring to this as *my* theory – partly that just means the theory I propose adopting in my argument here. Still, I am enthused and surprised by how well it works. We have all these intuitions that imply fairly specific connections between concepts, so deeply ingrained that we usually take them as given, though they can be threatened when we notice the world does not work in such a naïve way that we could take them all as primitives. Then, we can even question and reject every intuitive rationale for them, and yet, even after that, when we look at how they might be justified, not as primitives, but as part of how we need to live our lives with ourselves and others, we can get justifications for exactly the same connections between the concepts, based on that analysis and grounded ethical analysis instead of intuitions. Now, it may be misleading to say we get exactly the same thing. Parts of it have needed to be reinterpreted. But it was never unambiguously interpretable, certainly never unambiguously libertarian. Multiple interpretations of the same ideas have been available from the start.

It may also be more or less true that those intuitions and even instincts we seem to have that require certain things such as people being held responsible can be satisfied by (or while) explaining the concepts and their justification under this system. It may be that what would be more important for such a purpose is that those powerful intuitions are embraced and affirmed in the first place, and the reasoning given to it is secondary. Certainly there is nothing proving that only a libertarian conception can fulfil them for most people – though if we were to market a particular conception to them, it might work better to offer them that conception first instead of first offering them, say, a libertarian conception and then tearing that down, which might leave them with the feeling that *the* system of freedom and responsibility was torn down.

I do not actually know how workable this "solution" would be or how much a belief in ultimate origination or the like would be needed to keep people and society

functioning – any more than Smilansky[503] (6.2) or Strawson[504] (8.4) *knows*.[505]

9.5 The challenge of responsibility: A balancing act

I do not know whether thinking responsibility is based on metaphysics would make it easy to apply the concept in the real world, but it certainly does not make it easy to do so when it is based on something as vague as whether it is reasonable to hold someone responsible. I think this is a thing that is not supposed to be easy. Holding responsible – particularly in the sense of role responsibility, both moral and otherwise – is a two-edged sword that has the potential to do both harm and good.

On the good side, being held responsible may encourage a person to take charge of their life. It can guide people to consider the consequences of their actions to others, since they know they will be judged based on that. It can deter wrongdoing. Conversely, not holding someone responsible can lead them to be passive, to be heedless in their actions, or to commit wrong acts knowing that they can get away with them. Holding others or oneself responsible sends a powerful message, and it can be empowering on the one hand and just help to regulate unwanted behaviour on the other. This is why I think we cannot give up the concept of responsibility – though I have little doubt free-will sceptics who deny the existence of responsibility may be able to apply the same kind of thinking with other vocabulary.

However, all of this only works if it is indeed reasonable to ask someone to do the thing they are held responsible for. Though it is about applying a principle, it is about consequences too. If we hold someone responsible for something that they cannot change just by thinking of themselves as responsible and acting accordingly,

[503]Smilansky 2000.

[504]Strawson n.d.

[505]I could turn what I have said about how it may work into *a priori* arguments similar in to theirs – similar in terms of strength among other things – that would purport to prove that it does work, but I will not pretend that those arguments put so strongly would be sound.

the consequences are likely to be negative. The person can feel anxious, guilty, powerless, ashamed and so on. Consider, for example, holding people responsible for their sexual orientation, which they cannot in fact change, even though some people may see it as a choice. What is a person to do if they are held responsible for something like that? Since holding one party responsible often means not holding another party responsible, we may also miss out on holding the correct person responsible and thus miss the chance to change things for the better.

As a concrete example, one significant area where the challenge of responsibility comes up is in political decisions within a state about which policies to adopt to support the less well off. A balance needs to be struck here between supporting people and encouraging them to support themselves. More negatively, the balance needs to be found between the risks of letting people suffer from want or not reach their potential because they are not given a fair chance – and discouraging people from taking responsibility for themselves.[506] However, while there are arguments on both sides, discussion about this matter seems to be frequently muddled by the motivation to justify that those who already have power and money are deserving of it and should be allowed to keep it, leading to too easily attributing responsibility to those in a weaker position and belittling the difficulties they face.[507]

It is always possible to learn and form new principles about whom to hold

[506]See e.g. Schmidtz & Goodin 1998.

[507]See e.g. Kokko 2024b, O'Brien 2018, chapter 6, Piketty 2016 (e.g. pp. 376–378). Cf. Slattery 2014, p. 187, chapter 26, although, as usual, he frames his observations as a reason to advocate hard incompatibilism, seeing the belief in "free will" as a source of economic selfishness, mixed with a just-world belief. This is a somewhat valid way of putting it, but clearly, I suggest another. He also makes it sound as though the impossible kind of free will he calls such *would* make it acceptable to praise, blame, hold responsible and so on, as he constantly speaks of how it is not acceptable *because* there is no free will (see chapter 7, in this work for criticism on this assumption). After I pointed this out to him in an internet discussion, he made a point to say that libertarian free will would not be a *sufficient* condition for moral responsibility, but it is plausibly a *necessary* condition, which is closer to something I could agree with, though I would still go further in rejecting the supposed connection between libertarian free will and responsibility. Cf. also o*p. cit.*, 264.

responsible, and we already have such principles, some of which have been described above, such as the idea of disease as an excusing factor. However, since the world is always facing us with new situations, and the only real rule is that it is right to hold people responsible for what they can reasonably affect, it is not possible to formulate a full set of principles that would solve all situations. We just have to do our best in judging when to hold ourselves and others responsible and when not to – and also what form to do it in. The aim is to invoke the positive power of the concept of responsibility when possible, and avoid the negative effects it can have. It is always something of a gamble, but at least we can improve our odds at winning, even a great deal, by learning to be good at judging such situations.[508]

9.6 The challenge of freedom: Striving to become free

In the conception of free will that I have argued for, determinism is no threat to free will. There may be other threats, however; the definition of free will is substantial, and it is not something we have automatically. Though this work's main focus is on asking what we should even mean by "free will", in this section, I will briefly look at how much of it the evidence says we even have – and at what we should do when it turns out we do not have as much as we would like to.

9.6.1 Science does threaten freedom

For all that I have argued that determinism or the existence of scientific explanation does not in itself threaten free will, as I have defined it and as I have argued there is good reason to define it, it is still possible that what science more specifically tells us of ourselves shows that we do not have much of that kind of free will either. That is indeed the case: we are not the kind of rational animals we can easily imagine ourselves to be.

[508]The situation here is of the same form as when speaking about why higher-level laws contradict free will and reasons-responsiveness in 5.4.3.

Admittedly, if you read that scientists have disproven free will (or defended it), in my experience, the news is likely to be a red herring. Few examples show the importance of philosophical study and understanding better than the discussion around free will outside philosophy proper. It is typical for the conception of free will employed to be either so vague and as it were mysterious that it is easy to disprove it due to its internal problems, or to be just a naïve idea that free will equals indeterminism.

An example of a more sophisticated attempt that still fails is Robert Sapolsky's argument in his *Determined: Life without Free Will*.[509] Sapolsky presents evidence that the causes of human choices can be traced down partially on various different levels of description and different timescales, and that this means there is no room for free will, or at least not much[510]. He is ostensibly arguing mainly against compatibilists[511], and his evidence could make for a case against compatibilist free will on the basis that what determines human behaviour are the *wrong kind* of causes. However, he explicitly states his working definition of free will to be that some neuron in the process of decision-making seen on the neurological level should fire without a prior cause[512]. If *that* is the challenge to compatibilists, it makes no sense; compatibilists have little reason to want such an apparently indeterministic condition to obtain. You cannot challenge someone's view by asking them to prove something that is against the view. At the same time, Sapolsky does not explicitly say whether he thinks determinism is incompatible with free will or not, but he does rather late in the book argue that indeterminism definitely is incompatible with it (in a version

[509]Sapolsky 2023.

[510]Sapolsky deliberately stops short of saying there is no room whatsoever for "free will" so as not to overstate what he has proven (*Op.cit.*, p. 242.)

[511]*Op. cit.*, p. 11.

[512]*Op. cit.*, p. 15.

of the randomness argument)[513], without acknowledging how the indeterministic free will he opposes resembles his own definition of free will. Further, he equates "free will" with that which makes one morally responsible[514], but he does not address or note the questionableness of saying a random firing of a neuron makes you responsible. Thus, while Sapolsky's book provides much empirical knowledge for pondering our degrees of free will, his own attempt at studying the free-will question using this material stumbles due to a lack of grasp of the existing concepts and due to his not being able to keep track of his own philosophical argumentation.[515]

One discussion that looms large but which I have bypassed here entirely is the one starting with Benjamin Libet's experiments.[516] Shortly put, these experiments seemed to show that, when test subjects were making an arbitrary decision of when to press a button, a readiness potential in their brains indicated the decision before the moment they were consciously aware of it. This seems to imply that the impression of a conscious decision was an illusion, a powerless epiphenomenon following a decision made without consciousness. This presents a different kind of threat to free will than the one from determinism, as it might be seen as a different requirement of free will that the free decisions be made by our conscious part.

This needs to be mentioned even if only to explain why it is dismissed, but because it is a different argument than the one about determinism discussed in this

[513]*Op. cit.*, chapter 10.

[514]*Op. cit.*, pp. 12, 267.

[515]Ironically, Sapolsky shows a familiarity with many authors in the philosophical free-will discussion throughout the book, which does little to alleviate his confusion. Though I cannot assert this with much certainty, this might well be in part due to the way the randomness argument has been bypassed in so much of the discussion, creating a landscape more confused than it needs to be (cf. Kokko 2024c). In another part, it seems to be a stark demonstration of how philosophy is its own difficult field of expertise that is challenging to someone from another field – even when they understand their own field well and try to apply philosophy to it.

[516]Libet 1985.

dissertation, I will only treat it very briefly. Firstly, it should be noted that there are a number of different reasons why the Libet-style experiments might not show what they are supposed to be showing.[517] Secondly, regardless of Libet, science shows us a view of the world and of ourselves where we are not dualistic souls piloting separate material bodies, but where our consciousness is embodied and much of our cognitive functioning happens outside consciousness.[518] If we are this kind of beings, as it were essentially, we might want to give up on the view that only those decisions that come from our consciousness are ours. The non-conscious parts are a part part of ourselves too, and not such a radically different part. (This is related to the point about our being part of nature in section 9.1.2.) Instead of thinking *we* are not free, we can accept the independent reasons to think that the *I* has different boundaries than we might have thought, leading us to conclude the decisions still lie within *us*.

However, there is a different reason why our decisions being made unconsciously may be bad for freedom of will. When we see free will as centring around being responsive to the right reasons, the threat from unconscious influences that arises is that our unconscious processes may not be so responsive, and if they are not consciously accessible, we cannot even evaluate whether they are. Indeed, it seems in the light of empirical psychology that our conscious "reason" is often left to the role of riding atop unconscious processes and even downright delusionally rationalising them. Thus, these processes not only bypass conscious evaluation of reasons for choices but even distort it significantly. Further, the unconscious processes typically follow heuristics that, being heuristics, do not always suit the situation. They can be reasons-responsive, but they may not. Since these heuristics are at least in some part based on our evolutionary history in an environment different than the rapidly changing one we have built for ourselves during the last blink of

[517]See e.g. Neafsey 2021.

[518]See next paragraph and the references therein.

history's eye, they can even be ones that *usually do not work* in the present world.[519]

The way people choose how to vote in elections is an area rife with examples of lack of reasons-responsiveness, as it has been shown that the choice is affected by an ill-assorted heap of factors of which many have nothing to do with good *reasons* to choose a candidate or option.[520] Yuval Noah Harari mentions voting as one of the examples of a false belief in "free will" involved in liberal ideology and world view, where voters are assumed to make good, authentic decisions stemming from their own selves, whereas they are really affected by these other factors.[521] This is another example of the typical confusion of authors writing about free will: Harari assumes the presumed free will in question is of some libertarian sort, but the assumption that voters make good decisions for their own reasons would be better explained by saying that the voters are (falsely) assumed to be rational and reasons-responsive, not by assuming that they are undetermined or have an impossible combination[522] of control and indeterminism. Their being undetermined would compromise their

[519]For some overviews and perspectives on the psychology and neurology of behaviour, (ir)rationality, psychological heuristics, and the role of the unconscious, see Kahneman 2011, Montague 2008, Pinker 2021, Kahneman et al. 2021, Thaler & Sunstein 2009, and Sapolsky 2018.

[520]See e.g. Willman-Iivarinen 2023. It should be noted that choosing a candidate to vote for is also a matter that faces some external, objective difficulties. Insofar as some difficulty is completely objective or external, it is not a limitation of inner free will. For example, if one had to vote for one of two people, but there was no relevant information available about either at all, this would be an entirely external problem, and no amount of inner freedom would help with it.

All limitations are not purely inner or purely external. If it was hard to make a decision based on the available information due to the complexity of the situation, that would be an external limitation, but someone with the capacity to still make a good decision would have more inner capacity for freedom than someone who did not. In the light of this, it can be said that choosing a candidate also involves some difficulties that are more towards the external end of the spectrum. However, observation on what influences the choice in practice also shows plenty of factors that are so irrational that they can be counted as very much internal. (See the above source for details.)

[521]Harari 2018, pp. 45–47. Cf. Harari 2017, pp. 327–329.

[522]See 3.6 in the present work.

rationality in choosing,[523] and their being impossibly undetermined and in control instead of determined the right way is unnecessary.[524] Thus, Harari seems to present yet another example of a nominal incompatibilist wanting the consequences of the right kind of determinism[525] (or, if not personally desiring them, at least requiring them for free will).

There are also other challenges with respect to making the kind of decisions that follow our own best reasons and realise our interests. Some are related to the ability to keep making the right choices as part of a longer process that leads to a goal. I would say that it is possible to choose to lose weight – not merely to form an intention but carry it through – but that only highlights how a *choice* is not always an effortless, momentary mental act, but a difficult project. Further, we may be controlled by all kinds of external influences worming their way into our mental systems and causing disharmony. One way in which it may make sense to speak of this is that we can be influenced by "memes", ideas that are spread between people and selected for evolutionary fitness, potentially at the expense of their hosts.[526]

The list of things empirically threatening our ability to match the definition of free will I have proposed in this work is quite long. It would deserve its own study of corresponding length. Here, I will end with a simple observation: for empirical reasons, we are only partly free in terms of free will.

---

[523]See 3.3.4 in the present work.

[524]See 4.3 in the present work.

[525]Cf. 4.3.1 in the present work.

[526]A classic discussion of this topic is Blackmore 1999, though Blackmore seems to exaggerate the case. Instead of examining whether certain things can be explained in terms of memes, she seems to assume everything that might be is explained by them. (This comes on top of the limitation that the book is not based on careful empirical studies about the specific things it discusses, so it can mostly provide only interpretations and hypotheses in any case.) Thus, I invite the reader to trust their own judgement more than hers in reading this book.

9.6.2 Becoming the rational animal

It is not wonder that we used to think and even now think that the human is the rational animal. First, there is a strong element of truth to the idea. We do have some kind of ability to consider and decide things rationally. The problem is that this ability is very partial – even while it is rigged to make itself look much more complete than it is, making us the rationalising animal. Secondly, though, it would be a good thing if we were that. It would give us concrete control, and it would give us much of what we concretely want of free will, regardless of whether we persist on insisting that we also need to be unmoved movers somehow.

We find ourselves with only partial free will, with many limitations, many of which we may be only barely aware of. Yet, we can be aware of these limitations, and we can be aware of how to overcome them. We have some freedom to make ourselves more free, and we have every need to be so. We can also be said to have a responsibility to become as free as we can – both towards ourselves and others affected by our actions. This is what I want to introduce as the challenge of freedom, right here at the end: the challenge to become ever more free in the kind of sense that helps us live our lives better.

The idea of the human as the rational animal (or even non-animal) is dangerous if it lulls us into thinking we get to be rational like that for free. It is also simply wrong if it is taken to be true in a strong sense. Nevertheless, there is a place for it. I would view being the rational animal as part of the human "essence" as not a confident description of what we are, but as an ideal that we need to aim for. We are definitely by nature beings with interests who must pursue goals to advance those interests. In order to do this, we need to be free and in control of ourselves. The challenge of freedom, to become as free as we can, is a central aspect of human existence. The fact that we may not understand this very well only makes it more important to say it out loud. Even if what I am talking about was not "real" free will, and even if there were no other reasons, this is a reason why this kind of freedom is important.

List of References

Allison, H. E. (2012). *Essays on Kant*. Oxford: Oxford University Press.

Almeida, M., & Bernstein, M. (2011). Rollbacks, Endorsements, and Indeterminism. In R. Kane (Ed.), *The Oxford Handbook of Free Will (2nd edition)* (pp. 484–495). Oxford: Oxford University Press.

Almond, B. (1998). *Exploring Ethics: A Traveller's Tale*. Oxford ; Malden, Massachusetts: Blackwell Publishers Ltd.

Arendt, H. (1994). *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York ; London ; Ringwood ; Toronto ; Auckland: Penguin Books.

Balaguer, M. (2014). *Free Will*. Cambridge, Mass. ; London: The MIT Press.

Baumeister, R. F. (1996). *Evil: Inside Human Violence and Cruelty*. New York: Henry Holt and Company.

Bean, P. (1981). *Punishment*. Oxford: Martin Robertson.

Blackmore, S. (1999). *The Meme Machine*. Oxford: Oxford University Press.

Cabral, L. (2013). Are Economic Laws Compatible with Free Will? In A. Suarez & P. Adams (Eds.), *Is Science Compatible with Free Will: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience* (pp. 225–232). New York ; Heidelberg ; Dordrecht ; London: Springer.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: The University of Chicago Press.

Carroll, S. 2021. *AMA | March 2021* [Broadcast]. Retrieved 11 June 2025, from https://www.preposterousuniverse.com/podcast/2021/03/10/ama-march-2021/

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, D. J. (2006). *Strong and Weak Emergence*. https://consc.net/papers/emergence.pdf

Chalmers, D. J. (2012). *Constructing the World*. Oxford: Oxford University Press.

Chisholm, R. (2002). Human Freedom and the Self. In R. Kane (Ed.), *Free Will* (pp. 47–58). Oxford: Blackwell Publishers Ltd.

Cohen, J., & Stewart, I. (1994). *The Collapse of Chaos: Discovering Simplicity in a Complex World*. New York: Penguin Books.

Cyr, T., & Flummer, M. (n.d.). *Introduction to Debates about Free Will* [Broadcast]. Retrieved 28 May 2025, from https://thefreewillshow.com/episode-1/

Darwin, C. (2007). *The Descent of Man*. Darwin Online. http://darwin-online.org.uk/content/frameset?itemID=F955&viewtype=text&pageseq=1

Dawkins, R. (1999). *The Extended Phenotype: The Long Reach of the Gene*. Oxford: Oxford University Press.

Dawkins, R. (2016). *The Selfish Gene (40th anniversary edition)*. Oxford: Oxford University Press.

De Marco, G., & Cyr, T. W. (2024a). Manipulation Cases in Free Will and Moral Responsibility, Part 1: Cases and Arguments. *Philosophy Compass*, 19(12), e70009.

De Marco, G., & Cyr, T. W. (2024b). Manipulation Cases in Free Will and Moral Responsibility, Part 2: Manipulator-Focused Responses. *Philosophy Compass*, 19(12), e70008.

De Marco, G. (2025). Manipulation Cases in Free Will and Moral Responsibility, Part 3: Bypassing Responses. *Philosophy Compass*, 20(4), e70029.

Dennett, D. (2002). I Could not Have Done Otherwise—So What? In R. Kane (Ed.) *Free Will* (pp. 83–94). Oxford: Blackwell Publishers Ltd.
Dennett, D. (2003). *Freedom Evolves*. London: Penguin Books.
Dennett, D. C. (1973). Mechanism and Agency. In T. Honderich (Ed.), *Essays on Freedom of Action* (pp. 157–184). London ; Boston: Routledge & Kegan Paul.
Dennett, D. C. (1991). *Consciousness Explained*. London: Penguin Books.
Dennett, D. C. (2015). *Elbow Room: The Varieties of Free Will Worth Wanting. New Edition*. Cambridge, Massachusetts ; London: MIT Press.
Dennett, D. C., & Caruso, G. D. (2021). *Just Deserts: Debating Free Will*. Medford, MA: Polity.
Denyer, N. (1981). *Time, Action & Necessity: A Proof of Free Will*. London: Gerald Duckworth & Co. Ltd.
Dietrich, F., & List, C. (2016). Reason-Based Choice and Context-Dependence: An Explanatory Framework. *Economics and Philosophy*, *32*, 175–229.
Donagan, A. (1987). *Choice: An Essential Element in Human Action*. London ; New York: Routledge & Kegan Paul.
Doyle, R. O. (n.d.). Two-Stage Models for Free Will. *The Information Philosopher* website. http://www.informationphilosopher.com/freedom/two-stage_models.html
Doyle, R. O. (2013). The Two-Stage Model to the Problem of Free Will: How Behavioral Freedom in Lower Animals Has Evolved to Become Free Will in Humans and Higher Animals. In A. Suarez. And P. Adams (Eds.), *Is Science Compatible with Free Will: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience* (pp. 235–254). New York ; Heidelberg ; Dordrecht ; London: Springer.
Driver, J. (2022). The History of Utilitarianism. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy (Winter 2022 Edition)*. https://plato.stanford.edu/archives/win2022/entries/utilitarianism-history/
Duus-Otterström, G. (2007). *Punishment and Personal Responsibility*. Göteborg: Department of Political Science, Göteborg University.
Eagle, A. (2021). Chance versus Randomness. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Spring 2021 Edition)*. https://plato.stanford.edu/archives/spr2021/entries/chance-randomness/
Earman, J. (1986). *A Primer on Determinism*. Dordrecht: D. Reidel Publishing Company.
Ekstrom, L. W. (2011). Free Will Is Not a Mystery. In R. Kane (Ed.), *The Oxford Handbook of Free Will (2nd edition)* (pp. 366–280). Oxford: Oxford University Press.
Enqvist, K. (1998). *Olemisen porteilla*. Juva: Werner Söderström Osakeyhtiö.
Finch, A. (2020, August 17). *Episode 3: Logical Fatalism with Alicia Finch* [Interview]. https://www.buzzsprout.com/1244627/episodes/4945904
Fischer, J. M. (2011). Frankfurt-Type Examples and Semicompatibilism: New Work. In R. Kane (Ed.), *The Oxford Handbook of Free Will (second edition)* (pp. 243–265). Oxford: Oxford University Press.
Fischer, J. M., & Ravizza, M. (1993). Introduction. In *Perspectives on Moral Responsibility* (pp. 1–41). Ithaca ; London: Cornell University Press.
Fischer, J. M., & Ravizza, M. S. J. (2000). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge ; New York ; Melbourne ; Madrid: Cambridge University Press.
Floridi, L. (2011). *The Philosophy of Information*. Oxford: Oxford University Press.
Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *Journal of Philosophy*, *66*(23), 829–839.
Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5–20.
Gasparatou, R. (2008). What would you say then? The philosophical appeal to what one would say. *Sorites*, *21*, 63–70.
Gijsbers, V. (2023). The Paradox of Predictability. *Erkenntnis*, *2023*(88), 597–576.

Häggqvist, S. (1996). *Thought Experiments in Philosophy*. Stockholm: Almqvist & Wiksell International.
Haji, I. (2011). Obligation, Reason, and Frankfurt Examples. In R. Kane (Ed.), *The Oxford Handbook of Free Will (second edition)* (pp. 288–305). Oxford: Oxford University Press.
Harari, Y. N. (2017). *Homo Deus: A History of Tomorrow*. London: Vintage.
Harari, Y. N. (2018). *21 Lessons for the 21st Century*. London: Jonathan Cape.
Harris, S. (2013). *Vapaa tahto* (*Free Will,* 2012) (R. Mikkonen, Trans.). Kuopio: Scanria.
Harris, S. (2015). *Herääminen: Opas uskonnottomaan henkisyyteen* (*Waking Up*, 2014) (T. Kielinen, Trans.). Basam Books.
Hart, H. L. A. (2008). *Punishment and Responsibility: Essays in the Philosophy of Law. Second Edition*. Oxford: Oxford University Press.
Hartshorne, C. (1984). *Omnipotence: And Other Theological Mistakes*. Albany: State University of New York.
Hautamäki, A. (2018). *Näkökulmarelativismi: Tiedon suhteellisuuden ongelma*. Jyväskylä: SoPhi.
Hayward, M. K. (2019). Immoral realism. *Philosophical Studies*, *176*, 897–914.
Hobart, R. E. (1934). Free Will as Involving Determination and Inconceivable Without It. *Mind*, *1934*(169), 1–27.
Hodgson, D. (2002). Chess, Life, and Superlife. In R. Kane (Ed.), *Free Will* (pp. 249–256). Oxford: Blackwell Publishers Ltd.
Honderich, T. (1973). One Determinism. In T. Honderich (Ed.), *Essays on Freedom of Action* (pp. 187–215). London ; Boston: Routledge & Kegan Paul.
Honderich, T. (2002). *How Free Are You?: The Determinism Problem*. Oxford: Oxford University Press.
Honderich, T. (2011). Effects, Determinism, Neither Compatibilism nor Incompatibilism, Consciousness. In R. Kane (Ed.), *The Oxford Handbook of Free Will (second edition)* (pp. 442–456). Oxford: Oxford University Press.
Hurley, S. L. (2000). Is Responsibility Essentially Impossible? *Philosophical Studies*, *99*(2), 229–268.
Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Penguin Books.
Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgement*. London: William Collins.
Kane, R. (2002a). Free Will: New Directions for an Ancient Problem. In R. Kane (Ed.), *Free Will* (pp. 222–248). Oxford: Blackwell Publishers Ltd.
Kane, R. (2002b). Introduction. In R. Kane (Ed.), *Free Will* (pp. 1–26). Oxford: Blackwell Publishers.
Kane, R. (2011). Rethinking Free Will: New Perspectives on an Ancient Problem. In R. Kane (Ed.), *The Oxford Handbook of Free Will, second edition* (pp. 381–404). Oxford: Oxford University Press.
Kane, R. (2013). Can a Traditional Libertarian or Incompatibilist Free Will Be Reconciled with Modern Science? Steps Toward a Positive Answer. In A. Suarez & P. Adams (Eds.), *Is Science Compatible with Free Will: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience* (pp. 255–272). New York ; Heidelberg ; Dordrecht ; London: Springer.
Kane, R. (2017). Free Will, Bound and Unbound: Reflections on Shaun Nichols's bound [sic]. *Philosophical Studies*, *2017*(174), 2479–2488.
Kane, R., & Berofsky, B. (Eds.). (2011). Compatibilism without Frankfurt: Dispositional Analyses of Free Will. In R. Kane (Ed.), *The Oxford Handbook of Free Will (second edition)* (pp. 153–174). Oxford: Oxford University Press.
Kant, I. (2013). *Puhtaan järjen kritiikki* (*Kritik der Reinen Vernunft*, 1781/1787) (M. Nikkarla, K. Ranki, & O. Koistinen, Trans.). Helsinki: Gaudeamus Oy.
Kauffman, S. A. (2008). *Reinventing the Sacred: A New View of Science, Reason and Religion*. New York: Basic Books.
Kenny, A. (1973). Freedom, spontaneity and indifference. In T. Honderich (Ed.), *Essays on Freedom of Action* (pp. 87–104). London ; Boston: Routledge & Kegan Paul.

Knobe, J., & Nichols, S. (2011). Free Will and the Bounds of the Self. In R. Kane (Ed.), *The Oxford Handbook of Free Will, second edition* (pp. 530–554). Oxford: Oxford University Press.
Koi, P. (2024). Willpower as a metaphor. Published in D. Shoemaker, A. Amaya & M. Vargas (Eds.), *Oxford Studies in Agency and Responsibility Volume 8: Non-Ideal Agency and Responsibility*. Referenced here from https://philarchive.org/archive/KOIWAAv1
Koistinen, O. (2008). *Kant ja puhtaan järjen kritiikki*. Turku: Areopagus.
Kokko, V. (2023). How to Amend Christian List's Theory on Free Will to Answer the Challenge from Indeterminism. *Qeios*. https://doi.org/10.32388/73P1K2
Kokko, V. (2024c). Satunnaisuusargumentti ei jätä liikkumatilaa inkompatibilismille. *Ajatus*, *81*, 29–60.
Kokko, V. V. (2011). *Determinismi, indeterminismi ja tahdonvapaus*. Bachelor's thesis, university of Turku.
Kokko, V. V. (2014). *Todellisuuden tasot: Luonnon supervenienssi ja tieteenalojen väliset suhteet*. Master's thesis, University of Turku.
Kokko, V. V. (2018). Mitä 'hyvä' tarkoittaa? Arvottavien väitteiden mielekkyydestä. *Paatos*. http://www.paatos.fi/2018/07/16/mita-hyva-tarkoittaa-arvottavien-vaitteiden-mielekkyydesta/
Kokko, V. V. (2022). The Only Punishment. *After Dinner Conversation February 2022*, 59–73.
Kokko, V. V. (2024a). Write Down Every Moment: An Argument for Perspectivalism about A- and B-Theories of Time. In A. Carruth, H. Haanila, P. Pylkkänen, P. Telakivi (Eds.), *True Colors, Time after Time: Essays Honoring Valtteri Arstila* (pp. 187–203). Turku: University of Turku.
Kokko, V. V. (2024b, January 19). Oikeisto, vasemmisto ja vastuun haaste. *Turun Sanomat*.
Korman, D. Z. (2019). Debunking Arguments. *Philosophy Compass*, 14:12, e12638.
Kosko, B. (1993). *Sumea logiikka* (*Fuzzy Thinking: The New Science of Fuzzy Logic,* 1993) (K. Pietiläinen, Trans.). Helsinki: Art House Oy.
Kuula, K. (2006). *Helvetin historia: Pohjalta pohjalle Homeroksesta Manaajaan*. Helsinki: Kirjapaja.
Kuusi, O., & Virmajoki, V. (2022). Tulevaisuuksientutkimuksen filosofiset perusteet. In H. Aalto, K. Heikkilä, P. Keski-Pukkila, M. Mäki & M. Pöllänen (Eds.), *Tulevaisuudentutkimus tutuksi: Perusteita ja menetelmiä* (pp. 22–39). Turku: University of Turku.
La Caze, M. (2018). *The Analytic Imaginary*. Ithaca ; London: Cornell University Press.
Lanier, J. (2018). *Ten Arguments for Deleting Your Social Media Accounts Right Now*. London: The Bodley Head.
Libet, B. (1985). *Unconscious cerebral initiative and the role of conscious will in voluntary action*. *1985*(8), 529–566.
List, C. (2013). Free Will, Determinism, and the Possibility of Doing Otherwise. *Noûs*, *48:1*, 156–178.
List, C. (2019a). Levels: Descriptive, Explanatory, and Ontological. *Noûs*, *53:4*, 852–883.
List, C. (2019b). *Why Free Will Is Real*. Cambridge, Mas.: Harvard University Press.
List, C., & Rabinowicz, W. (2014). Two Intuitions about Free Will: Alternative Possibilities and Intentional Endorsement. *Philosophical Perspectives*, *2014*(28), 155–172.
Loughrey, D. (1998). Second-Order Desire Accounts of Autonomy. *International Journal of Philosophical Studies*, *6*(2), 211–229.
Mackie, J. L. (1965). Causes and Conditions. *American Philosophical Quarterly*, *12*, 245–265.
McGinn, C. (1994). *Problems in Philosophy: The Limits of Inquiry*. Oxford ; Cambridge, Mass.: Blackwell.
McKenna, M. (2011). Contemporary Compatibilism: Mesh Theories and Reasons-Responsive Theories. In R. Kane (Ed.), *The Oxford Handbook of Free Will (second edition)* (pp. 175–198). Oxford: Oxford University Press.
Mele, A. R. (2002). Autonomy, Self-Control and Weakness of Will. In R. Kane (Ed.), *The Oxford Handbook of Free Will (1st edition)* (pp. 529–548). Oxford: Oxford University Press.
Messadié, G. (1996). *A History of the Devil* (*Histoire Générale du Diable*, 1993). M. Romano (Trans.). New York: Kodansha America.

Midgley, M. (1993). The origin of ethics. In P. Singer (Ed.) *A Companion to Ethics* (pp. 29–43). Malden, MA ; Oxford ; Carlton, Victoria: John Wiley & Sons, Incorporated.
Midgley, M. (1994). *The Ethical Primate: Humans, Freedom and Morality*. Malden, MA ; Oxford ; Carlton, Victoria: Routledge.
Montague, R. (2008). *Miksi valita tämä kirja? Miten teemme päätöksiä* (*Why Choose this Book?*, 2006) (K. Pietiläinen, Trans.). Helsinki: Terra Cognita.
Moore, D. (2022). Libertarian Free Will and the Physical Indeterminism Luck Objection. *Philosophia*, *50*, 159–182.
Nadelhoffer, T., Murray, S., & Dykhuis, E. (2023). Intuitions About Free Will and the Failure to Comprehend Determinism. *Erkenntnis*, *88 (6)*, 2515–2536.
Nagel, E. (1971). *The Structure of Science: Problems in the Logic of Scientific Explanation*. London: Routledge & Kegan Paul.
Nagel, T. (1986). *The View from Nowhere*. Oxford: Oxford University Press.
Nagel, T. (2012). *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*. Oxford: Oxford University Press.
Neafsey, E. J. (2021). Conscious intention and human action: Review of the rise and fall of the readiness potential and Libet's clock. *Consciousness and Cognition*, *94*.
Nichols, S. (2006). Folk Intuitions on Free Will. *Journal of Cognition and Culture*, *6*(1–2), 57–86.
Niiniluoto, I. (2007). Probabilistinen kausaliteetti. In H. Gylling, I. Niiniluoto, E. Vilkko (Eds.), *Syy* (pp. 221–236). Helsinki: Gaudeamus.
O'Brien, J. (2018). *How to Be Right: In a World Gone Wrong*. London: WH Allen.
O'Connor, T. (2011). Agent-Causal Theories of Freedom. In R. Kane (Ed.), *The Oxford Handbook of Free Will (2nd edition)* (pp. 309–328). Oxford: Oxford University Press.
Ofstad, H. (1967). Recent Work on the Free-Will Problem. *American Philosophical Quarterly*, *4*(3), 180–207.
Pereboom, D. (2011). Free-Will Skepticism and Meaning in Life. In R. Kane (Ed.), *The Oxford Handbook of Free Will (2nd edition)* (pp. 407–424). Oxford: Oxford University Press.
Peroutka, D. (2022). Partial Compatibilism: Free Will in the Light of Moral Experience. *Organon F*, *29*(1), 2–25.
Piketty, T. (2016). *Pääoma 2000-luvulla* (*Le capital au XXI siècle*, 2013) (M. Ollila & M. Tillman-Leino, Trans.). Helsinki: Into Kustannus Oy.
Pink, T. (2005). Will, the. In E. Graig (Ed.), *The Shorter Routledge Encyclopedia of Philosophy* (pp. 1055–1056). London ; New York: Routledge.
Pink, T. (2011). Freedom and Action without Causation: Noncausal Theories of Freedom and Purposive Agency. In R. Kane (Ed.), *The Oxford Handbook of Free Will (2nd edition)* (pp. 349–355). Oxford: Oxford University Press.
Pinker, S. (2011). *The Better Angels of Our Nature: The Decline of Violence in History and Its Causes*. Longon: Allen Lane.
Pinker, S. (2021). *Rationality*. London: Penguin Books.
Prosser, S. (2013). The Passage of Time. In H, Dyke, A. Bardon (Eds.), *A Companion to the Philosophy of Time* (1st ed., pp. 315–327). Chichester: Wiley-Blackwell.
Pust, J. (2024). Intuition. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/intuition/
Reid, T. (2010). *Essays on the Active Powers of Man* (K. Haakonssen & J. A. Harris, Eds.). Edinburgh: Edinburgh University Press.
Sapolsky, R. (2018). *Behave: The Biology of Humans at Our Best and Worst*. London: Vintage.
Sapolsky, R. (2023). *Determined: Life without Free Will*. Dublin: The Bodley Head.
Schlossberger, E. (1992). *Moral Responsibility and Persons*. Philadelphia: Temple University Press.
Schmidtz, D., & Goodin, R. E. (1998). *Social Welfare and Individual Responsibility*. Cambridge: Cambridge University Press.

Searle, J. R. (2007). *Freedom and Neurobiology: Reflections on Free Will, Language, and Political Power*. New York: Columbia University Press.
Shapiro, L. A. (2000). Multiple Realizations. *The Journal of Philosophy*, *197(12)*, 635–654.
Singer, P. (n.d.). All Animals Are Equal (retrieved from https://spot.colorado.edu/~heathwoo/phil1200,Spr07/singer.pdf and originally appearing in T. Regan & P. Singer (Eds.) 1989: Animal Rights and Human Obligations, Cambridge: Cambridge University Press, pp. 148-162.)
Slattery, 'Trick. (2014). *Breaking the Free Will Illusion for the Betterment of Humankind*. (Location not indicated): Working Matter.
Smilansky, S. (2000). *Free Will and Illusion*. Oxford: Clarendon Press.
Smilansky, S. (2011). Free Will, Dualism, and the Centrality of Illusion. In R. Kane (Ed.), *The Oxford Handbook of Free Will (2nd edition)* (pp. 425–441). Oxford: Oxford University Press.
Solomon, A. (2002). *Keskipäivän demoni: Masennuksen atlas* (*The Noonday Demon: An Atlas of Depression*, 2001.) (A. Schroderus, Trans.). Helsinki: Tammi.
Speak, D. (2011). The Consequence Argument Revisited. In R. Kane (Ed.), *The Oxford Handbook of Free Will (second edition)* (pp. 115–130). Oxford: Oxford University Press.
Stanovich, K. (2003). *How to Think Straight about Psychology*. Boston: Pearson Education, Inc.
Stewart, I., & Cohen, J. (1997). *Figments of Reality: The Evolution of the Curious Mind*. Cambridge: Cambridge University Press.
Stockton, F. (1882). The Lady, or the Tiger? *The Century Magazine*, *25*(1), 83–86.
Strawson, P. F. (n.d.). Freedom and Resentment. *The Determinism and Freedom Philosophy Website*. https://www.ucl.ac.uk/~uctytho/dfwstrawson1.htm
Taylor, C., & Dennett, D. (2011). Who's Still Afraid of Determinism? Rethinking Causes and Possibilities. In R. Kane (Ed.), *The Oxford Handbook of Free Will, second edition* (pp. 221–240). Oxford: Oxford University Press.
Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving Decisions about Health, Wealth and Happiness*. London: Penguin Books.
The 2020 PhilPapers Survey. (2020). https://survey2020.philpeople.org/
Tononi, G. (2013). On the Irreducibility of Consciousness and Its Relevance to Free Will. In A. Suarez & P. Adams (Eds.), *Is Science Compatible with Free Will: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience* (pp. 147–176). New York ; Heidelberg ; Dordrecht ; London: Springer.
Ushpiz, A. (2016, October 12). The Grossly Misunderstood 'Banality of Evil' Theory. *Haaretz*. https://www.haaretz.com/israel-news/2016-10-12/ty-article/the-grossly-misunderstood-banality-of-evil-theory/0000017f-db5b-d3a5-af7f-fbffca7e0000
van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.
van Inwagen, P. (2000). Free Will Remains a Mystery. *Philosophical Perspectives*, *14*, 1–20.
van Inwagen, P. (2002). The Mystery of Metaphysical Freedom. In R. Kane (Ed.), *Free Will* (pp. 191–195). Oxford: Blakcwell Publishers Ltd.
van Inwagen, P. (2011). A Promising Argument. In R. Kane (Ed.), *The Oxford Handbook of Free Will (2nd edition)* (pp. 475–483). Oxford: Oxford University Press.
Vargas, M. (2011). Revisionist Accounts of Free Will: Origins, Varieties and Challenges. In R. Kane (Ed.), *The Oxford Hadbook of Free Will, second edition* (pp. 457–474). Oxford: Oxford University Press.
Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.
Vincent, N. A. (2011). A Structured Taxonomy of Responsibility Concepts. In N. A. Vincent, I. van de Poel, & J. van den Hoven (Eds.), *Moral Responsibility: Beyond Free Will and Determinism* (pp. 15–35). Dorcrecht ; Heidelberg ; London ; New York: Springer.
Visala, A. (2018). *Vapaan tahdon filosofia*. Helsinki: Gaudeamus Oy.
Watson, G. (1987). Free Action and Free Will. *Mind*, *1987*(Vol 96, No. 382), 145–172.

Watson, G. (1993). Responsibility and the Limits of Evil: Variations on a Strawsonian Theme. In J. M. Fischer & M. Ravizza (Eds.), *Perspectives on Moral Responsibility* (pp. 119–148). Ithaca ; London: Cornell University Press.

Watts, A. W. (1966). *The Book: On the Taboo Against Knowing Who You Are*. New York: Vintage.

Wiggins, D. (1973). Towards a Reasonable Libertarianism. In T. Honderich (Ed.), *Essays on Freedom of Action* (pp. 31–61). London ; Boston: Routledge & Kegan Paul.

Williamson, T. (2020). *The Philosophical Method: A Very Short Introduction*. Oxford: Oxford University Press.

Willman-Iivarinen, H. (2023). *Satunnainen äänestäjä*. Helsinki: Miratio.

Willmot, C. (2016). *Biological Determinism, Free Will and Moral Responsibility: Insights from Genetics and Neuroscience*. (Location not indicated.) Springer Nature.

Wise, S. M. (2005). Animal Rights, One Step at a Time. In C. R. Sunstein & M. C. Nussbaum (Eds.): *Animal Rights: Current Debates and New Directions* (pp. 19–50). Oxford: Oxford University Press.

Wolf, S. (1990). *Freedom within Reason*. New York ; Oxford: Oxford University Press.

Wolf, S. (2002). Sanity and the Metaphysics of Responsibility. In R. Kane (Ed.), *Free Will* (pp. 145–163). Oxford: Blackwell Publishers Ltd.

Wong, D. (1993). Relativism. In P. Singer (Ed.), *A Companion to Ethics* [ebook version]. Malden, MA ; Oxford ; Carlton, Victoria: John Wiley & Sons, Incorporated.

Yanofsky, N. S. (2019). *Perustellun tiedon ulkorajat: Mitä tiede, matematiikka ja logiikka eivät voi kertoa* (*The Outer Limits of Reason*, 2023) (T. Perhoniemi, Trans.). Helsinki: Terra Cognita.

Yli-Hemminki, E., Melander, S., & Nuotio, K. (2022). Johdatus rangaistusteorioihin: Miksi rangaista ja millä perusteilla? In E. Yli-Hemminki, S. Melander, K. Nuotio (Eds.), *Rikoksen ja rangaistuksen filosofia* (pp. 21–36). Helsinki: Gaudeamus.

Young, R. (1993). The implications of determinism. In P. Singer (Ed.), *A Companion to Ethics* [ebook version] (pp. 748–760). Malden, MA ; Oxford ; Carlton, Victoria: John Wiley & Sons, Incorporated.

Appendices

10 Appendix A: Levels of reality and description

The concept of levels of description is used throughout this work, often in conjunction with (in)determinism. I introduce it briefly in the main text in section 2.1.2. In this section, I dive into it more deeply, both to characterise and explain what it is about and to give more precise definitions. Thus, this appendix tries to answer any possible questions about what I really mean by levels of description. The last section contains more general reflection on how to understand the relationships between different levels, and explanation in general.

Before going into the definition proper, I look at the related concepts of *scrutability* and *emergence* by the way of illustration.[527]

10.1 Scrutability

The concept of *scrutability* was introduced by David Chalmers in his book

[527]For some related ideas of how the world is viewed through different viewpoints, though not conceptualised so much as levels that are figuratively atop one another, see Hautamäki 2018; Nagel 1986; and La Caze 2018, chapter 5, which takes a critical view of the just mentioned Nagel 1986.

Constructing the World.[528] It is not very significant in this work, but it relates to levels of description, and it will be useful to refer to its sometimes, so I will explain it briefly. The idea is that a class of truths *B* is scrutable from a class of truths *A* iff, if an agent knew all truths of type *A*, then the agent could know all truths of class *B* based on that.[529] As an example, it might be that if a godlike thinker similar to Laplace's demon (see section 2.1.1 in the present work) knew all physical, phenomenal and indexical truths about the world, as well as knowing that it is enough to know all these things (the "that's all" clause), then it could know every truth of every sort there is to be known about the world.[530]

Since we are talking in terms of levels of description, we can say the question of scrutability is whether something on one level is scrutable from another level, or whether a whole level is scrutable from another level.

10.2 Emergence

Phenomena on "higher" levels of description are often said to be *emergent* from those on lower levels. Entire levels might also be described as emergent. The basic idea of emergence is that the emergent thing is somehow novel or unexpected compared to the level it emerges from, the *emergence base*.

A strangely rarely noticed aspect of emergence is that it can be either synchronic – with the emergent level existing at the same time as the emergence base – or diachronic – with the emergent phenomenon being something that arises over time from the emergence base, such as life evolving from lifeless matter.[531] It is odd

[528]Chalmers 2012. For a summary of the concept by myself in Finnish, see Kokko 2014, chapter 2.

[529]Chalmers 2012, e.g. pp. xiii-xiv.

[530]This idea is discussed in various places throughout Chalmers 2012. For a summary and more precise references for the locations in Chalmers's book, see Kokko 2014, pp. 10–11.

[531]See Kokko 2014, pp. 47–49.

that this is not noticed more often, because logically these two senses of *emergence* could be quite different. However, as far as I have observed, this potential confusion rarely causes any practical confusion. In the present work, I am almost entirely concerned with synchronic emergence. Levels of description mainly relate to synchronic emergence, and only indirectly to diachronic.[532]

A related term is "supervenience". In some sense, it means much the same thing as *emergence*, although without the diachronic aspect. Each term has multiple different meanings, and the difference between them has been explained in more than one way, or they may be treated as being different without a proper justification.[533] However, if one summarises the general idea of the range of meanings of either term (again ignoring the diachronic version of emergence), one ends up with the same rough idea. I will use *emergence* as my default term of choice in this dissertation.[534]

A major question related to emergence and levels of description is whether the different levels are in fact derivable from each other or at least some ultimate level (see 10.6), or whether new phenomena emerge on a higher level in a way that cannot be explained even in principle by looking at the lower level – basically the same thing as whether they are scrutable from it (10.1), although I will not commit to

---

[532]Indirectly as in: a series of events over time that is related to diachronic emergence, such as evolution, may be described on different levels of description, and it may be necessary to describe their dynamics on an emergent level to understand them at all (see Kokko 2014, chapter 5). These levels are still synchronic with each other, though, since even though the events they describe have duration in time, it is these same events at the same moments that are described on the different levels.

[533]See for example how Chalmers defines "supervenience" in 1996, chapter 2 and "emergence" in 2006 without in either case acknowledging that he is talking of at least roughly the same thing each time, and further that his "logical supervenience" corresponds to his "weak emergence" and his "nomological supervenience" to his "strong emergence".

[534]I made the opposite choice in my Master's thesis Kokko 2014, so what I said of "supervenienssi" (the Finnish version of "supervenience") there might as well be read as applying to what is called "emergence" here.

saying it is exactly the same thing. Emergence always means the emergent level is novel compared to the lower one in *some* sense, but if it is not derivable from the lower level even in principle, then it is *strong emergence*; otherwise, *weak emergence*.[535] Temperature on the macroscale is weakly emergent in relation to the microscale of individual molecules and their movements because it can be explained how the movements of the particles causes the phenomena associated with temperature.[536] If, instead, no such explanation was possible – which might even mean that there would be nothing going on at the microscale that correlates with temperature – then temperature would be strongly emergent.

That all said, I am here going to assume levels of description are related via weak emergence. This means I have to explain their relationships and build an overall coherent system. Strong emergence is logically possible, and perhaps in some limited sense it has to be part of the laws of nature, but as an explanation, it only posits a brute fact, and since I can do without that, I will.

10.3 Levels of description

When we describe or think about the world – or anything at all, if there are things that are not part of "the world" – we always use some kind of a conceptual system, the "language" in a broad sense in which the description or thinking is done. This could mean very many different kinds of things, and our actual descriptions and thoughts virtually always use a vague conceptual system rather than a well-defined one. Nevertheless, our ordinary conceptual schemes approximate better-defined levels of description, and to make philosophical arguments using the concept of levels of description, I need to make it more precise. In any case, the topic here is not really the ways people talk about the world but the way in which the world is

[535]See e.g. Chalmers 2006.

[536]See e.g. Enqvist 1998, pp. 67–68 for a non-technical characterisation of this idea and a bit of philosophical musing on it.

possibly *correctly* described, so we want to talk about accurate and unambiguous descriptions of it.

Luciano Floridi[537] builds a whole methodology of philosophy based on *levels of abstraction*, which is a version of the same concept that I refer to here as *levels of description*.

To start off defining levels of description less precisely but possibly more informatively, a particular level of description is something that talks about particular kinds of entities, particular objects and kinds of qualities of those objects, particular kinds of events and processes.[538] On some microphysical level, we will speak of individual particles and properties that they have. On a macrophysical level, we will talk about material objects and the kinds of properties they have, without mentioning that they are made up of particles. We will talk about different phenomena on a psychological, or societal, or economic level. Each level has its own kind of vocabulary and entities, cats or quarks, and statements about observational entities on each level can be confirmed by the kind of observations, more or less direct or indirect, that relate to how those entities are defined.[539]

From here onwards, I am roughly following Floridi's definition and discussion of levels of abstraction.[540]

The objects, properties, processes and so on on a level of abstraction can be

[537]Floridi 2011.

[538]Cf. Taylor & Dennett 2011, pp. 233–235.

[539]A good explanation and some interesting use of the idea is found in List 2019b; see the quotation in the present work in 2.1.2. From List on this, see also List 2019a.

[540]As part of his project of defining what is a level of abstraction, Floridi provides more precise and specific definitions than I do here, and I do not wish to commit to saying exactly the same things as he does. I have no need to be that precise, and it might even be a hindrance, since the concept of levels of description I use here is very general and includes all kinds of different levels. My explanation is heavily inspired by his definitions, however. Below, I will quote parts of his definitions in the footnotes to indicate where they *roughly* correspond to what I am saying.

presented as a universe made up of a set of states of affairs. For example, if we have the entity *C*, which is a particular cat, and *M*, which is a particular mat, then we might have the state of affairs that "*C* is on *M*" be true, or not. (This is obviously a description that abstracts away where *C* is on *M.* It is also unable to handle borderline cases where *C* is kind of on *M* and kind of not, unless of course we give the predicate a very precise definition, in which case it is prone to excluding some borderline cases, which in turn could be problematic.) *C* might also be black, or not, and it might be eating at time *T* or not.

C might also weight exactly 4,111 kilograms, or not. However, if we start describing numerical valuables as on-off states like this, we will have infinitely many of them for each number (or at least as many as can be conceptualised on that particular level of abstraction). Instead, we can describe the composition of the world as described on a particular level of abstraction by saying that it consists of the values of a number of variables, with each variable having a range of possible values; for *C* being on *M*, it is *yes* or *no* and possibly something else like *kind of* if we can handle that[541], but for *C's* weight, its value is *4,111 kg* – instead of the weight of *C* being replaced by a whole group of variables like "Does *C* weight 4,112 kg?" Naturally, this can also be applied to other kinds of predicates with multiple options to make things simpler, such as "colour of *C*" = *black*.[542]

Levels of description can contain variables for all kinds of truths. All of the examples above are observational truths. When speaking of observational truths, some can be seen as complex, consisting of more simple truths, but some must be

[541]As in fuzzy logic, on which see e.g. Kosko 1993.

[542]The idea in this paragraph is roughly the same as the following definition by Floridi:

> A *typed variable* is a uniquely named conceptual entity (the *variable*) and a set, called its *type*, consisting of all the values that the entity may take. Two typed variables are regarded as *equal* if and only if their variables have the same name and their types are equal as sets. A variable that cannot be assigned well-defined values is said to constitute an *ill-typed variable*[.]
> (Floridi 2011, p. 48, italics in original.)

regarded as basic. For example, perhaps we can see *C* in some particular part of space, and *M* in some particular part of space, and derive the fact that *C* is on *M* from the relationship of these two... but perhaps, at the same time, our visual observation of where *C* is and the other one about where *M* is are not analysed further. It would be hard or impossible to say in practice where we are drawing the lines on a particular occasion of what is a basic observation not derived from others – just thinking about it might cause us to start reducing the location of *C* to the location of those body parts of it that we can see, for example – but, again, to define a level of description precisely, we need to do this. In any case, some observations must be regarded as basic in order for others to be constructed out of them.

So, with observational truths, we derive their truth from particular observations, which are the values of variables on the level of description.[543] There are other kinds of truths, although we do not need them much when talking about freedom and determinism. The most relevant other class is what might roughly be called logical and mathematical truths – those that are true in a way derivable from their definitions and relationships, without need to refer to the empirical world.[544] Normative truths are a more complicated case that are analysed briefly elsewhere (8.3.2).

A point worth noting here briefly is that the way I am putting this reveals a way in which "different" levels of description, with different sets of variables, can be equivalent. Reality itself does not need to be any different if one level of

[543]This sentence and the previous paragraph relate to the idea behind this definition provided by Floridi:

> An *observable* is an interpreted typed variable, that is, a typed variable together with a statement of what feature of the system being modelled it represents. Two observables are regarded as *equal* if and only if their typed variables are equal, they model the same feature and, in that context, one takes a given value if and only if the other does. (Floridi 2011, p. 48, italics in original.)

[544]Anything that is either analytic or *a priori*; the further distinctions are of no relevance here, since I am only mentioning this class of truths in a vague generalisation and will hardly use it for anything in the present work.

description says that "*C* is black" = true than if another says that "colour of *C*" = *black*. Both could be confirmed with the same observation. There is an important caveat, however: levels of description are *in principle* incommensurable. To define what is the same observation on different levels, you need to have a third level containing both of the two and rules for relating them. Producing such a third level is perfectly doable in principle and often in practice, but one must be conscious of it. If you compare two levels of description without being aware you are doing that, you are acting as if you can just compare things to reality itself on no particular level. This may be fine much of the time, but since you are not actually using the only possible way to describe things, you should be aware of that in case it becomes relevant. Relationships between levels will be discussed shortly in section 10.4.

So, we might say that what a level of description is in principle is a set of variables, each defined by its possible values and the conditions by which those values are set (e.g. what colour a cat looks to be or – to keep things really simple here – what an electronic scale's display says when you weigh it). Let us call this a *level of description in the narrow sense*.[545] It is immediately noticeable that this means that we are in practice changing our levels of description all the time. If we meet a new cat, we effectively add a new variable for its colour and so on.[546] So,

[545]The level of description in the narrow sense is analogous to Floridi's definition of a level of abstraction:

> A *level of abstraction (LoA)* is a finite but non-empty set of observables. No order is assigned to the observables, which are expected to be the building blocks in a theory characterized by their very definition. An LoA is called *discrete* (respectively *analogue*) if and only if all its observables are discrete are discrete (respectively analogue); otherwise it is called *hybrid*.
> (Floridi 2011, p. 52, italics in original.)

[546]We could in principle – certainly not in practice – define a level of description so that it contains as basic variables the possible values of all most primitive observations we can make (if we knew what they are) – the smallest possible patch of colour in some part of our visual field, say – and derive all our possible observations from that, and on this basis be ready to describe every possible observational state of affairs without needing to introduce new variables. We could also do something similar without basing the basic variables on the actually most basic possible observations, but just decide to draw the line somewhere, and then keep on describing the world only in

what we are really talking about when[547] talking about levels of description are the *kinds* of variables available. You can add new variables and new states of affairs without changing the level of description by definition, as long as you keep them within some basic categories, say, the colour of an object. We can call this a *level of description in the broad sense*, but from now on (and in other chapters), whenever I just say "level of description" without further modification, the broad sense is what I mean. That is also what was meant when characterising a level of description in more general terms in the previous sections, such as the level(s) where there are only elemental particles (but not only specific particles $p_1...p_n$).

Obviously, people conversing in natural language seldom restrict themselves to a single, clearly defined level of description. In many cases, jumping between levels, and/or merging them, comes very naturally indeed. This is shown indirectly in one of Floridi's examples of a level of abstraction:

> [T]o describe the state of a traffic light in Rome one might decide to consider an observable *colour* of type {*red*, *amber*, *green*} that corresponds to the colour indicated by the light. This option abstracts the length of the time which the particular colour has been displayed, the brightness of the light, the height of the traffic light, and so on.[548]

The point I am raising with this is that it is quite natural for Floridi to have to list things that are abstracted away – things that a human can easily observe are also

these terms, in which case we might miss things but could still in principle describe everything in that level of description.
Of course, if we used such a reductive level of description, we would not be able to refer to such things as the identity of cat *C*; we would instead have to reduce it to empirically observable things like spatial continuity, if we referred to it at all.

[547]This is true everywhere in the present work except parts of the present section, and it probably applies to every other work referenced here except Floridi 2011.

[548]Floridi 201, p. 51. Italics in original.

part of the situation, even though the person may also well ignore them when just looking at what colour the light is showing.

Floridi writes that this level of abstraction abstracts away certain things, which is true enough, but it is also true it abstracts them away only in comparison with some broader level of description/abstraction in which they are included. We can say that it abstracts them away in comparison with reality itself, of course, but every time we describe reality in such a way as to include that which was described away, we are creating a new level of description; we cannot get to reality itself except through one such level or another.

Vagueness about levels of description can become a problem if something about the question at hand depends on the level of description chosen. That is something that I aim to avoid in this work by paying close attention to the different levels I discuss.[549]

10.4 Relationships between levels

As I stated in section 10.3, the relationship between levels of description can only be understood by articulating it on a level of description that contains all of the levels of description whose relationship is to be explained together with the explanation of that relationship. Since different levels of description are about the same underlying reality, there are unsurprisingly many cases in which different levels have some kind of a correspondence relationship. To explain what the relationship is, we need to

[549]Something I do not discuss at this point more closely is change of states of affairs as described on a level of description. For those interested, here is a definition by Floridi related to it:

> The *behaviour* of a system, at a given LoA, is defined to consist of a predicate whose free variables are observables at that LoA. The substitutions of values for observables at that LoA. The substitutions of values for observables that make the predicate true are called *system behaviours*. A *moderated LoA* is defined to consist of an LoA together with a behaviour at that LoA.
> (Floridi 2011, p. 53, italics in original.)

Appendix B in the present work adds some discussion on this topic.

explain how the variables on one level correlate with those on another.[550] To explain why and how there is such a relationship, we need to know something about what is really going on in the world to relate these levels. Such questions will mostly be answered in the next section (10.5), but this section begins to introduce the idea. The basic idea is usually that more complex phenomena, and levels of description describing them, are built "atop" simpler levels.

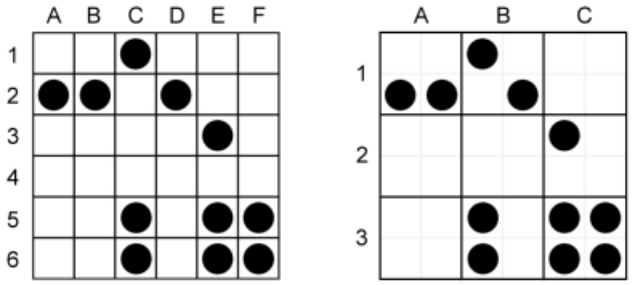

Figure 1: Two levels. L1 on the left, L2 on the right

I will represent levels of description here with simple diagrams. Both the lower level L_1 and the higher level L_2 represent a description of the same[551] system S with the following limitations: each consists only of a grid of squares, which are fixed, and the possible states of the system on the level consist of each square either containing or not containing a circle (at least one, and the number is not counted beyond that).

As we can see in Figure 1, L_1 has six times six squares (36), and L_2 only three

[550]See Floridi's explanation of a gradient of abstractions in Floridi 2011, pp. 54–58. His formal definition for a gradient of abstraction is on page 55, but it is more complicated than the previous ones I have quoted, and I do not think repeating it here out of context would be helpful.

[551]Of course, the sameness can only be defined on a third level of abstraction that connects L_1 and L_2.

times three (nine). Now consider the contents of the squares in this image: in this state of the system, on L_1, squares A-2, B-2, C-1, C-5, C-6, D-2, D-5, D-6, E-5, E-6, F-5 and F-6 have circles in them. Meanwhile, L_2 cannot give exactly the same information about S, even if it is describing the same system, because it contains a differently sized grid. We can see that in this image, squares A-1, B-1, B-3, C-2, and C-3 contain circles.[552]

As a matter of fact, these images *can* be seen as conveying as much of the same information as possible given the limitations of the levels that they are describing. This is explained by the concepts of *multiple realisability* and *coarse-graining*.

10.5 Multiple realisability and coarse-graining

As I state in the section about emergence (10.2), I will assume levels of description, when they can be arranged in a "hierarchy", relate to each other through weak emergence. This means that something novel emerges on the higher level, but nothing can really be added to the lower level to make it so. It turns out that the major way in which this is achieved, both in principle and actually in nature, is by *subtracting* information instead: the higher-level description contains different concepts than the lower-level one in such a way that knowing everything the higher-level concepts does not allow one to know – even in principle – everything about the correct lower-level description.[553] Scrutability (10.1) works only upwards, not downwards.

This is based on what is called *multiple realisability* in philosophy. If L_1 is the

[552]Square names in L_1 and L_2 do not refer to the "same" squares in any sense when their names are the same, e.g. "A-1" when speaking of L_1 is not the same as "A-1" when speaking of L_2.

[553]Christian List speaks of this general thing as "levels of grain" (List 2019a, pp. 856–857), though he speaks of something similar even when he is not putting it under that heading (e.g. *op. cit.*, 858).

lower level and L_2 is multiply realised on L_1, that means a complete description of the system at hand in terms of L_2 can correspond to multiple different descriptions of the system in terms of L_1. The term *coarse-graining* is a broadly equivalent term from the natural sciences: L_2 is coarse-grained with respect to L_1 when L_2 is multiply realisable on L_1. A more coarse-grained level has a lower *resolution*: like a low-resolution picture, it cannot show equally small details at the less coarse-grained description with better resolution.

I demonstrate the idea with diagrams again:

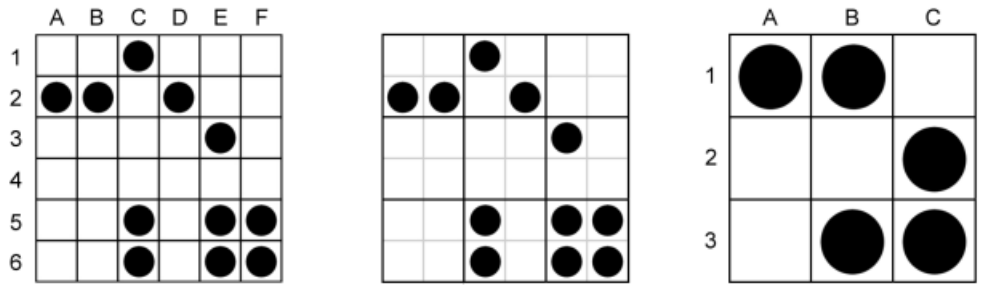

Figure 2: Resolution and coarse-graining

Consider Figure 2. It shows a simple transformation (or translation) from a state of S on L_1 to a state on L_2. The rule for this transformation is as follows: for each square of L_2, consider the four squares that would overlap it if the grids were laid atop one another; A-1 on L_2 corresponds to A-1, A-2, B-1, and B-2 on L_1, and so on.[554] Then fill in the squares on L_2 as follows: any time there is a circle in any of the squares on L_1 corresponding to a given square on L_2, that square gets a circle on it on L_2. Otherwise, it is empty.[555]

[554]This could be explained more rigorously, but I see this as the clearest explanation, besides the illustration itself.

[555]The rule could also be, e.g., that the bigger square gets filled when there are at least two filled smaller squares corresponding to it. This would reflect a situation in which a single filled smaller square is too small to be noticed on the higher level and lower

Now, consider Figure 3. This image shows a state of the system *S* on level L_1 that is different from the one in Figure 2 – yet following the same rule for moving from L_1 to L_2, we get the *same* state on L_2.

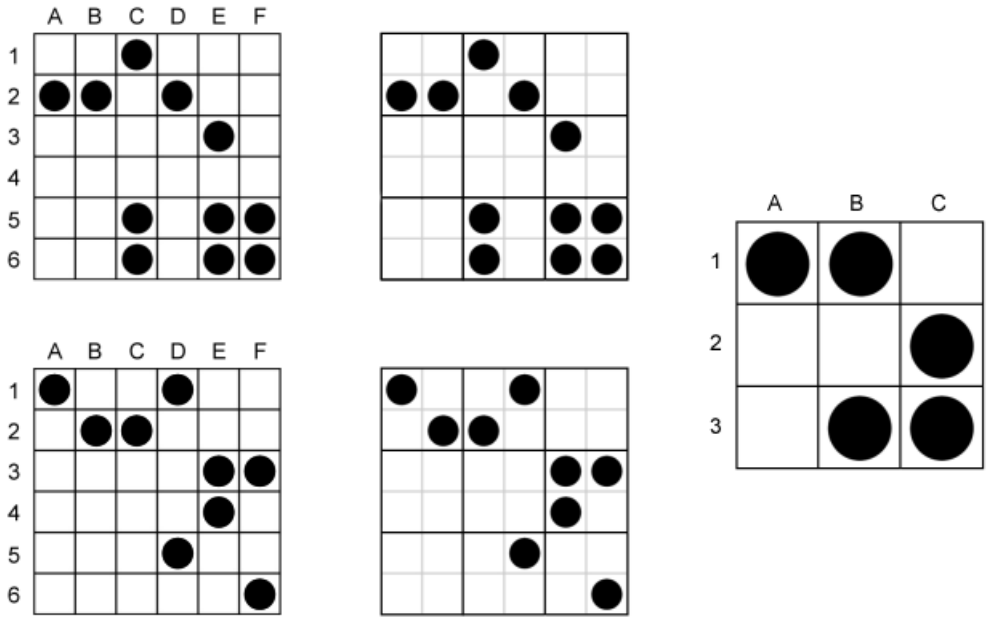

Figure 3: Multiple realisability

This is a demonstration of the concept of multiple realisability, which is generally defined as follows:

A level of description L_2 is multiply realisable in terms of another level of description L_1 iff

- both can be used to describe the same system *S*, and
- it is possible that when *S* is described in terms of L_1 as being in state St_1, and

resolution.

> when S is described in L_1 as being in another state St_2, it is nevertheless in both cases described in terms of L_2 as being in the same state St_3.[556]

A straightforward example of multiple realisability in nature is coarse-graining in physics. In the example mentioned before, the temperature of a gas follows from the average kinetic energy of the individual gas molecules; knowing the temperature of a volume of gas, you lose sight of the individual molecules and their movements on which it is based.

The importance of different and higher levels of description lies in that, as List points out in the quote before (section 2.1.2), lower-level descriptions are all trees and no forest[557]. Higher-level descriptions abstract away large amounts of detail that is confusing, mere noise in the context where the higher-level description is useful. We have no possibility of detecting or processing the information about the movements of every particle of air in a room, but we can detect the temperature – and be affected by it. The macroscopic phenomenon of temperature can even cause the macroscopic physical phenomenon of a fire starting.[558] Such a thing as evolution by natural selection is also a higher-level pattern that explains what happens in its purview in a way that cannot be captured on the lower level. Nature itself works on different levels.[559] Thus, to lose information in multiple realisation is also to find

[556]Lawrence Shapiro (2000) has proposed a criterion for real multiple realisability that would exclude such a broad definition. See Kokko 2014, pp. 34–36 for an argument for why such a criterion is unwarranted.

[557]As in the saying "Not seeing the forest for the trees."

[558]I am not concerned here with questions of whether higher-level phenomena really have causal power. The point is that they have explanatory power at least. A quick note on this is that since a cause is often analysed as being something that is only a relevant part of the causal story leading up to the effect (see e.g. Mackie 1965), this kind of causal power seems quite compatible with higher-level phenomena having causal power. However, I will not claim this is a decisive argument in the complex debate.

[559]Further, higher-level phenomena can be multiply realisable in such a sense that the same kind of phenomenon can appear in completely different bases, not just be realised differently in the same kind of physical base. For example, evolution by natural

new ways of understanding phenomena thanks to the elimination of noise.

In terms of the free-will question, it is obvious that we are confronted with at least two levels, an ultimate or microphysical level and a psychological level. A neurological level might be a third one in between. What multiple realisability might mean between levels relevant for free will is explored among other places in 3.4.3, 5.4 and 9.1.3.

As a final note on this topic, multiple realisability and coarse-graining are such universal concepts that they even apply to philosophical concepts and distinctions. Thus, in 3.4.2, I state that there is an important sense in which agent-causal and non-causal theories of agency are the same, even though they are rivals, and that I can treat them as being the same for the purposes of my ongoing argument. This is a valid move insofar as there really is no *relevant* difference in the current context, meaning that a view that coarse-grains the two concepts into one preserves that which matters.

10.6 The ultimate level

Physicists speak of and search for a hypothetical "Grand Unified Theory" or "theory of everything" that would be the ultimate, fundamental physical theory from which everything else can be derived in principle. A theory unifying quantum mechanics and general relativity is regarded as the likely candidate at this moment. There is criticism that it is misleading to talk of such a theory as a "theory of everything", since it would not explain everything in practice, and though I would say it would be unsurprising if such a theory were to be found, there is no guarantee that it exists.[560] Nevertheless, such a theory would be fundamental in some sense – though perhaps in no other sense than what directly follows from its definition.

---

selection, a familiar phenomenon from biology, could happen inside a computer program.

[560]For discussion on these points, see Cohen & Stewart 1994, pp. 364–365 and Stewart & Cohen 1997, pp. 41–43.

If you knew this ultimate physical theory, and you knew the full description of the universe in terms of this theory – say, if the theory described everything in terms of certain fundamental particles, and you knew the state of every such particle in every sense mentioned in the theory – you would in some sense know the entire state of the universe. Everything would be scrutable from this level (see section 10.1, in the present work).[561] Every other description of the universe would somehow have to be built on this level.

The worry about determinism is usually seen as a worry that it applies on the ultimate level: that the universe is "really", at the bottom, deterministic. If there is some way of describing what happens as deterministic, that suffices for the basic incompatibilist worry. (However, see 2.6.1.)

The ultimate level does not, in principle, need to be tied to a theory in physics. You could believe in an ultimate level existing without affirming that physics is all there is. For example, if you believed there existed separately both mind and matter, irreducible to each other, the ultimate level could contain a full description of both of these components. (See also 2.1.6.) It will be defined in terms of levels of description above.

This ultimate level as here defined should not be confused with the sense in which Saul Smilansky uses the expression "ultimate level". (See 6.2 as well as 5.3.4 and 5.3.3 in the present work.) I consistently refer to that other concept as "Smilansky's ultimate level".

[561]Thus, compared to the example in that section, you would not need the separate class of phenomenal truths, and you would only need the indexical truths and the "that's all" clause to make certain kinds of statements that are in a sense only necessary for knowing everything as a technicality.

10.7 Parts, wholes and homunculi

> Leaving something out is not a feature of failed explanations, but of successful explanations.
> -Daniel Dennett[562]

When there are multiple different levels of description that can be placed in a hierarchical gradient, the lower-level ones and the higher-level ones do not contain the same terms. The higher levels do not show all the detail of the lower levels, so moving to a higher level will cause the loss of some terms from the lower level. It also works in the other direction: moving down to a lower level, as when performing a reductive explanation or analysis of the higher level, will cause some of the terms and entities from the higher level to vanish out of sight.[563] If a higher level of description contains a whole that is analysed into its parts on the lower level, the lower level will contain mention of the parts as entities, not the whole as an indivisible entity.

This may seem trivial, but what has significant consequences for the discussion of many philosophical issues is that it is, evidently, often difficult to apply this realisation in practice. This problem takes the form of people thinking that an explanation on the lower level needs to introduce the terms from the higher level as entities on the lower level as well, or it has not explained them.

Consider, for example, this description of a problem allegedly encountered by one model of free will[564]:

---

[562]Dennett 1991, p. 454.

[563]For more concrete examples of what this means, you can refer back to the description by Christian List in 2.1.2 in the present work.

[564]The model is the one by Robert Kane (2002a, 2011), discussed in 3.4.7 in the present work. Its details are currently not relevant, though I should note that, even though I am disagreeing with a criticism of it here, I disagree with the view itself for other reasons. Note also that the criticism I am presenting here is unrelated to the one I present when discussing Kane's theory.

> One way to criticise Kane's view of agency is to ask what is left in it for the agent themselves to do. According to Kane, the person taking action consists of states of mind (such as emotions, beliefs, desires and goals). The agent's actions are results of these states of mind. In other words, the agent's decision is caused by their goals and other states of mind[.] … Now, we may ask where the agent themselves is in all this: if the agent is nothing but a collection of states of mind, how does the agent themselves take part in decision-making?[565]

In other words, if all of an agent's actions are caused by the agent's states of mind, and the agent is nothing but their states of mind, what part does the agent actually take in anything? To distil the point even further to show what is wrong with it: If that which is the agent is that which causes their own actions, what is the agent's role in all this?[566]

Though this makes little sense said the way I did here, intuitively, it seems to have a meaning: when we speak of parts of the agent, we no longer speak of the agent as something separate from them. Visala goes on to speak of agent causality as an answer to this alleged problem. Agent causality is the idea that the agent, as a substance or entity of some sort, causes the agent's choices or actions, as opposed to only events causing other events[567]. Further discussion of agent causality in this work is found in subsection 3.4.2[568], but right now, the basic idea is that it introduces the

---

Visala's book (2018) that I am referring to here is an introduction to the philosophical questions of free will and determinism, so it is not implied the author is expressing his own opinion in this passage.

[565]Visala 2018, p. 121. My translation.

[566]Cf. Knobe & Nichols 2011 for how such intuitions have been studied.

[567]Visala 2018, pp. 122–127.

[568]In addition, somewhat as with Kane's theory above, agent causality is discussed there in a different sense than the one in which it is mentioned here. Here, it is mentioned as being what amounts to a homunculus (see below), whereas in the other location, it is

agent as an undivided whole. States and processes within the agent do not cause the agent's actions; the agent does, and we are not to ask what the agent is made up of. Thus, the agent is a homunculus, a little person within the person, who is making the choices in a way that may not be analysed.

The reason that it seems intuitive, natural or necessary to "explain" things in this fashion probably has to do with out psychological tendencies: when we think in terms of agency, we do not naturally think in terms of reducing the agent to parts. This topic is discussed in section 5.3.

It is the nature of levels of description that they contain different entities, even when they are related to each other in a gradient. It is the nature of explanation that an explanation should not simply refer back to the thing explained. It is the nature of analysis that the thing analysed is "taken apart".

A story illustrating this point is that of the pastor and the steam engine, recounted by R. E. Hobart while discussing free will and attributed by him to someone called Paulsen:

> We have been accustomed to think of a thing or a person as a whole, not as a combination of parts. We have been accustomed to think of its activities as the way in which, as a whole, it naturally and obviously behaves. It is a new, an unfamiliar and an awkward act on the mind's part to consider it, not as one thing acting in its natural manner, but as a system of parts that work together in a complicated process. Analysis often seems at first to have taken away the individuality of the thing, its unity, the impression of the familiar identity.
>
> For a simple mind this is strikingly true of the analysis of a complicated machine. The reader may recall Paulsen's ever significant story about the introduction of the railway into Germany. When it reached the village of a certain enlightened pastor, he took his people to where a locomotive engine was standing, and in the clearest words explained of what parts it consisted and how it worked. He was much pleased by their eager nods of intelligence as he proceeded. But on his finishing they

discussed as an attempted answer to the randomness argument.

> said : “Yes. yes, Herr Pastor, but there’s a horse inside, isn’t there?” They could not realise the analysis. They were wanting in the analytical imagination. Why not? They had never been trained to it. It is in the first instance a great effort to think of all the parts working together to produce the simple result that the engine glides down the track. It is easy to think of a horse inside doing all the work. A horse is a familiar totality that does familiar things. They could no better have grasped the physiological analysis of a horse’s movements had it been set forth to them.[569]

The horse in a steam engine is a kind of homunculus. Hobart’s claim that we are accustomed to think of familiar wholes as such is part of a line of thought that I discuss especially in 5.3.

In this work, I do not accept any kind of homunculus explanations. They are only an invitation to stop explaining. If we ask about some aspect of how our decision-making capacities work in greater detail, for example, it is not an answer to this question to say anything that implies a small decider inside the decision-making process. (See also 3.3.7.) This is especially relevant when we are asking about whether the process is deterministic or indeterministic and what follows from that (see especially chapters 3 and 4). Most things can be analysed back to their constituents on a different level of description, and once this is done, we no longer have the original whole on that level. This does not mean that we have either disproven or neglected its existence, merely that we have actually analysed and explained something about it.[570]

[569]Hobart 1934, p. 3; the space before the colon is in the original, though it would be impossible to mark that understandably with the usual notation for such things.

[570]See Harris 2015, chapter 2 for a solid – and utterly fascinating – argument against the unity of the conscious mind or the existence of a simple, unified “soul” behind it. For another similar line of thought, see Slattery 2014, chapter 15. This line of argument also relates to the discussion of the nature of the self in the present work, in sections 8.3 and 9.6.1.

11 Appendix B: A definition for determinism

As with levels of description, I briefly introduce the definition of "determinism" that I use in my arguments throughout this work in section 2.1. In this appendix, I define it more precisely, discuss its implications for the debate further, and examine an illuminating example – the ideas of Thomas Reid – to show what it means and does not mean.

This appendix assumes that the reader has read Appendix A to know my full definition of levels of description, though the short version in section 2.1.2 is enough to have a rough idea.

11.1 Universal determinism

What I need my definition of "determinism" to do is to say that, given a certain state of affairs at a certain moment T_1, only one total state of affairs will be possible at T_2. In order to say this, I will need to specify what I mean by "possible" in this context. I will keep the answer simple, and that does not leave many options. Logical possibility is too strong. Nomological possibility seems about right: we want to talk about determinism in terms of how the world works. Nomological possibility means being possible within the limits set by the laws of nature. Determinism on this level is broadly what is called *causal determinism*. That said, though *nomological possibility* and *causal determinism* give a good idea of what kind of determinism I am aiming at defining (at least to those already familiar with them), neither term is exactly right. I am not tying my definition to laws of nature – instead, I am tying it to whatever rules can be said to govern the world on the level of description at hand.

This will become clear when I introduce the role of levels of description in defining determinism and the broader equivalent of nomological possibility I do use, below – in 11.3 and 11.5, respectively.[571]

I do need to refer to the kind of possibility I will be talking about with some term, though, even before I go into detail about it. I will call it *nomological possibility**.

Note that logical possibility is implied by nomological possibility*. If something is logically impossible, it is nomologically impossible* as well, even if its impossibility follows from something else than the laws. It is still true that it is not possible under the current "laws".

Figure 4 below illustrates the basic idea of determinism and indeterminism in the same kind of manner as I represented levels of description above (10, p.306). On the left is the deterministic scenario, in which only one future state can follow. On the left is the indeterministic scenario in which more than one state can follow. The point of illustrating the idea with such an image will become clear when I talk about determinism and levels of description below (11.3), as I will do more complicated things with similar images there.

[571]For different formulations of determinism to similar effect, see Denyer 1981, p. 6, Earman 1986, p. 13, van Inwagen 1983, p. 65, Wiggins 1973, p. 36, Balaguer 2014, p. 13, specifically referring to actions: Honderich 2002, p. 63.

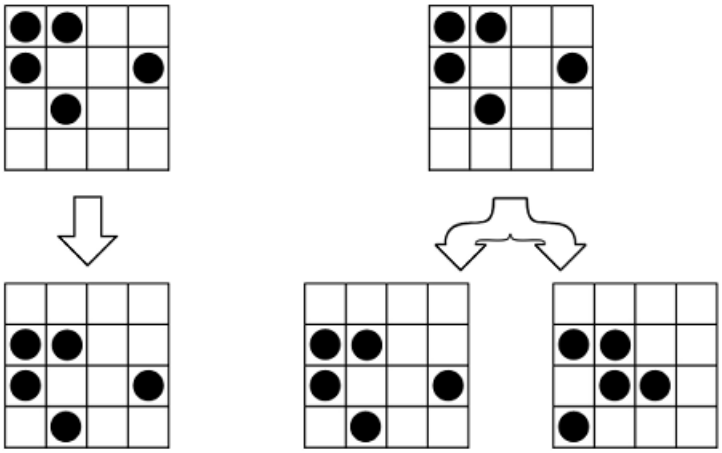

Figure 4: Determinism and indeterminism

Thus, on this first level of defining determinism, we can present the following abridged definition. This is a definition for universal determinism since we have not yet introduced any notion of what it would mean for only a part of the universe to be deterministic.

> The universe is deterministic if and only if, given any total state of the universe at time T_1 and any later[572] time T_2, only one total state of the universe is nomologically possible* at T_2.

[572]It would be possible to define a kind of determinism where, in addition to only one later state being possible as defined here, given any *later* state of the system, only one *earlier* state at a given time would be possible. This requirement seems to be stricter than what is usually thought of as determinism, and in any case, it seems to have no bearing on the question I am discussing, so I will not take it into account in my definitions of determinism or any of the forthcoming discussing. Cf. Slattery 2014, chapter 16.

11.2 Local determinism

Besides knowing what it means for the entire universe to be deterministic, we will want to define what it means for only some limited part of it to be so. The "part" can be defined in multiple ways. It could mean a continuous area of space during a particular continuous stretch of time, but it could refer to an individual object or process, too. I see no evident problems in leaving open the possibility of coming up with different notions of what can be seen as a part of the universe in the sense intended in this definition. However, there will be a more precise definition given below (11.4) after the use of levels of description in defining determinism has been introduced. For now, let us simply look at what it means for determinism to apply to only a part of the universe, whatever that part may be.

Because it is possible for a part of the universe to be affected by the rest of the universe, we cannot define determinism in a part of the universe in the same way as in the whole universe, i.e. that certain conditions within the part at T_1 would lead to only one possible set of conditions within the part at T_2.[573] If that were the case, the part would really be separate from the rest of the universe. It might be able to affect the rest of the universe, but the rest of the universe could not affect it in any way.[574] After all, if it could, then the conditions within the part would not be sufficient to determine what happens in it, because you would have to know about the rest of the universe as well. This would also mean the rest of the universe could not even be observed from within the part, because that would make the observers within the part have different observations, which, if the rest of the universe was indeterministic,

[573]Cf. Earman 1986, p. 33.

[574]This might actually be useful if the "rest of the universe" were a prediction machine, as discussed in section 5.2.2, and the "part" would be the universe being made predictions about. This would prevent the recursivity that becomes a problem for prediction in the thought experiment. However, saying that it would be useful does not mean that it would be justified. It would also not be useful to do *in this work*, as these "problems" are not a problem in my proposed theory, merely a problem for something we would hypothetically doing as part of the thought experiment.

would then make different futures possible within the part.

We do want such observations to be possible in our deterministic part of the universe. It should be part of the universe in such a strong sense that it can be affected by the rest of it. Therefore, the way I will define local determinism is that given the same state of affairs at time T_1 *and the same subsequent input from outside* P *up to and including* T_2, only one state of affairs is nomologically possible* within the area at T_2. The rest of the universe may give indeterministically inspired input, but the deterministic area will react to it in a lawful manner.

If the universe is partly deterministic and partly indeterministic, there must be some limits of some sort that the indeterminism cannot cross. Maybe only some behaviours of some particles are indeterministic but statistical, maybe indeterministic exceptions to the laws of nature only occur at certain (small) probabilities. This requirement could also be fulfilled by indeterminism only occurring within human choices.

The universe must be lawful enough to contain repeatedly observable rules, properties and objects, or else nothing could be known about it and in any case there could be no life or structure in it in the first place.[575]

11.3 Determinism on levels of description

So far, it has been assumed that when we speak of determinism being true, it is true on the ultimate level (10.6). This would mean that the world really is like that at the bottom; it is, as it were, not a matter of perspective. However, in a different sense, it would still be true to say that it is a matter of perspective. It could still be that, if we look at the world under some other level of description, we can truly and meaningfully say that determinism does not hold on that level. Conversely, it could be true that determinism holds on some higher level of description even if it does not

[575]Cf. Slattery 2014, chapter 20, and p. 328, note 21 (where "acausal" means indeterministic).

on the ultimate level.[576]

To give the basic idea of determinism as applied to levels of description:

> Determinism holds on a level of description *L* iff, given any total state of the system, *as described in the terms available in* L, and any input *I as described in the terms available in* L, only one total state *as described in the terms available in* L is possible at T_2.[577, 578]

Thus, if we are talking about (in)determinism on some level L_1, differences that may be recognisable in some other level of description L_2 do not count when identifying the same initial state, final state, or input. Recall that two different ultimate-state descriptions of the universe might correspond to the same higher-level description on some level L_1 but count as two different states on some other level L_2.

Emergence of higher levels from lower ones as weak emergence based on coarse-graining can explain both the case where there is determinism on the lower level and indeterminism on the higher as well as the opposite case of indeterminism

[576]Compare this with List's idea of possible world on different levels of description (List 2019a, pp. 858–859).

[577]This definition is noticeably incomplete because I am saving the complete version for a later subsection.

[578]What if there are things in the universe that could be described on some level L_x, and which (as seen on a level incorporating both *L* and L_x) can have effects on things described on *L*, but which are not themselves describable on *L*? In the present context, there are two possibilities for how this goes. Firstly, it like leads to indeterminism on *L* because the changes brought about by the undetectable entities cannot be accounted for. Secondly, it *might* be possible to introduce rules for the behaviour of entities on *L* that take into account the effects of the undetectable L_x entities – the L_x entities would still be undetectable, but their effects would be described as rules about the behaviour of things on *L*. The latter option could only be possible in very limited cases – such as taking it as a given that things will fall towards the ground, even though the Earth and its gravity are something not described on *L* – or it might require building the L_x entities into the rules of *L* in such a way that it would just be the equivalent of introducing those entities to *L* after all, even if they would only be presented through variables affecting the behaviour of explicitly recognised *L* entities.

on the lower level and determinism on the higher.

Determinism on the lower level and indeterminism on the higher level: If L_2 sees two possible scenarios as identical in terms of both initial state and input but L_1 sees a difference in either initial state or input between the two possible scenarios, and both L_1 and L_2 see two possible final states, then the system is proven as indeterministic at L_2 but not at L_1, because only at L_2 do the same conditions lead to two possible different outcomes. (See Figure 5.)

Indeterminism on the lower level and determinism on the higher level: Conversely, if the initial state is the same according to both L_1 and L_2, and so is the input from the rest of the universe. However, it is nomologically possible* for the system to have what are in (the lower) L_1 recognised as two different states, but in L_2 count as the same state. In this case, the system is deterministic at L_2 but indeterministic at L_1, as well as at the fundamental level. (See Figure 6.)

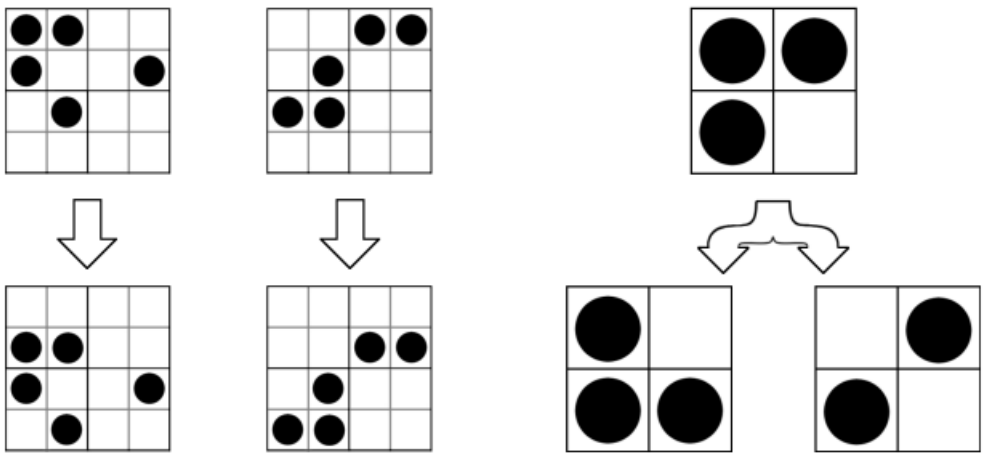

Figure 5: Determinism on the lower level, indeterminism on the higher level

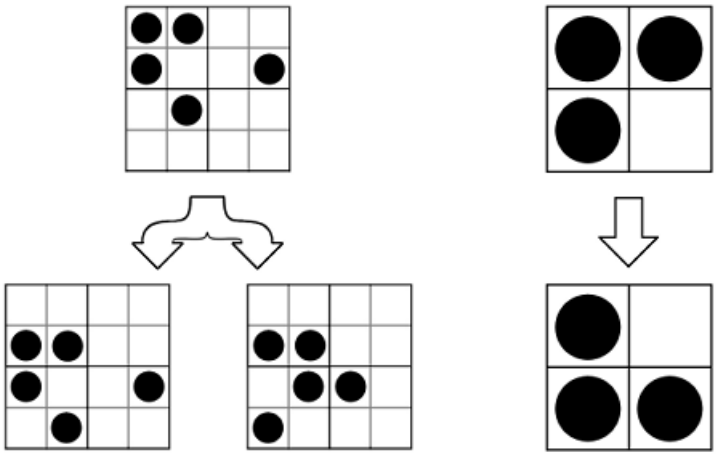

Figure 6: Indeterminism on the lower level, determinism on the higher level

Note that while this picture gets some of the idea across, it is inadequate in a way explained shortly in 11.5.

Chaotic systems are a good example of determinism on the lower level but indeterminism on the higher level. Their development over time is so sensitively dependent on initial conditions that indefinitely small changes in the conditions on the lowest relevant level will cause differences so large that they will show up on the higher level. Thus, no coarse-grained higher-level description can be used to create accurate predictions even as described on a higher level, except over time spans too short for the differences to manifest. This is why weather can be predicted, but the predictions keep getting less accurate the further in the future they are.

While it is possible for indeterminism to exist on a lower level while determinism exists on a higher level in the same system, this is conceptually less easily achieved than the opposite case, since the lower level is the one carrying more information. A system does not need to be chaotic for small differences on the lower level to accumulate until they have enough of an impact on the higher level. Something needs to limit the indeterminism on the lower level enough for it to stay

within such limits that the upper level can stay deterministic. Decoherence in quantum mechanics has been presented as such a factor[579], though I will not try to explain it here. Luís Cabral[580] argues that another case of indeterminism on a lower level coupled with determinism on a higher level can be found in the relationship between the indeterminism of the microeconomic level of people's individual decisions and the determinism of the macroeconomic level of the large-scale phenomena of economics.

11.4 Defining parts of the universe via levels of description

Levels of description gives us a way to define "part of the universe" more precisely, though still very generally. For the most part, it should not matter too much how we define what can be a part. I see no evident way in which it would make trouble for our current discussion if we understand the part as being any of a number of possible things – areas, processes, or objects, at least.[581, 582]

However, it is easy enough to use the notion of levels of description again to say what could be defined as a part. A part of the universe on a given level of description can be defined as either the set of certain individual variables, or any and all variables (including potentially news ones) fulfilling a certain condition such as appearing within a certain spatial area.

---

[579]Cohen & Stewart 1994, p. 271.

[580]Cabral 2013.

[581]One thing I do *not* want to say that the part is a particular law of nature or other dynamic rule. (*Dynamic rule* will be defined next in 11.5.) It can make sense to say that, but that goes too far beyond what we are doing here otherwise. It is better expressed in present terms like this: is part *P* of the universe is affected by both deterministic and indeterministic dynamic rules, then *P* is indeterministic.

[582]Cf. Kokko 2014, p. 77-78.

11.5 Change, possibility, and levels of description

In this subsection, we come to the question of how to understand what has so far been called "nomological possibility*". I have not yet explained what kinds of "laws of nature" might apply on different levels of description, but it should be clear enough from my examples that not all levels of description have what we normally call "laws of nature".

The solution to this problem is that I will next explain what is the rough equivalent of laws of nature on an arbitrary level of description and, having done that, I will explain how the kind of possibility meant in the definition is tied to this concept rather than actual laws of nature.

The basic idea is simple: besides of values for various variables, a level of description can also contain rules for how those values can change. For example, in Figure 6 above, every circle on L_1 moves down one square in one time step (or, to be more precise, for every square, if and only if that square had a circle above it, it will now have a circle in it). Of course, there may not be minimal steps in time (or space or other values of variables) like above – the rules of change for a level may be such that they tell what happens at any arbitrary moment after the first one.

Such rules can be deterministic by prescribing only one possible state at T_2 given one state at T_1, or they may give several options. When deterministic, they must have some generalisability to be meaningfully[583] different from indeterminism. If you simply had a set of rules saying that from each exact state of the whole system follows some other state, and there would be no other connection between the first state and the second, then unless the system was so small that it kept repeating the same exact states, the rules would not be detectable and the system might as well be indeterministic. Examples of generalised rules would be "all massive objects exert gravitational force (of a strength calculable in a specific way) on each other" or, in

[583]Incompatibilists might hold that this kind of determinism still contradicts freedom, making the distinction meaningful in their eyes, because this kind of determinism still disallows alternative possibilities in the individual case (see 2.4.2).

the above picture, "all circles move one step down on each time step." These can both be stated in a form that is applicable to several different situations, in fact all possible situations.[584]

Indeterministic rules for change in the system on a given level of description may also be statistical – giving a fixed probability for each different outcome to happen. Of course, this strictly speaking also applies to indeterministic rules that are not statistical, since the chance that a particular outcome happens must have some probability of happening in principle. Presumably if someone says something is just indeterministic without being statistical, this means that the different possibilities are all equally likely.

If we look at Figure 6 again, we can notice the upper level of description is not actually covered by any deterministic rule. It is merely coincidence that the two different changes on L_1 amount to the same change on L_2. Nothing guarantees this, and further, L_2 could be even more random than L_1 for the same reason it is in the previous Figure 5. In order for L_2 to have a deterministic rule for change, an indeterministic L_1 would have to be restricted in its indeterminism in a way that would relate suitably to phenomena on L_2. It seems that in such a case as this, L_1 would have to be "aware" of the bigger squares on L_2, so that the circles on L_1 would only move in ways detectable on L_2. Though it is possible to formulate rules that relate an indeterministic lower level to a higher level so that the latter becomes deterministic, to represent them with such images gets complicated and artificial.[585]

---

[584] "Gravitational force" can be defined as something that combines with all other forces affecting a body's movement, which makes this "law" true regardless of what other forces there are – but then it must only be regarded as part of the totality of rules that affect bodies in that system. In other words, if we present gravity as exceptionless like this, then unless we are talking about a world in which there is no other force than gravity, the gravitational "law" is not yet a rule that tells what will happen, because other forces can also have an effect on the final outcome.

[585] Here is one such rule for the kind of set of squares that we have been using as an example, although we can take this grid to be indefinitely large:

On L_1, there are vertical columns two squares wide, each of which can contain only one circle. For each T_n and T_{n+1}, a circle moves either two steps straight down, or two steps down and one to the side, though note that it still cannot cross over to

In the real world, perhaps the primary way this is achieved by nature is through the statistical effects of coarse-graining. For example, there are so many more ways for a volume of gas to expand to fill the entirety of a space than not to do so that a higher-level rule predicting it will do so would never be wrong in practice even if the behaviour of the individual molecules would be truly random.

In any case, different levels of description can have quite different rules for how the system can change. They can be deterministic, or they can be statistical, random, or just vague. They can be expressed with exact mathematics or inexact words. Alternatively, a level of description might lack them and only have the conceptual vocabulary to describe single states of the system, in which case determinism and indeterminism are not an issue on that level. There is no need nor possibility to define the rules of change for a level of description more precisely than this.

What matters for defining determinism is that possibility in the definition is taken to mean possibility according to the rules that govern the change of the system on that level of description. This is now the explanation for what has been called nomological possibility*. It is not exactly nomological possibility, since all of those rules are hardly laws of nature,[586] but it is the broad analogue for laws of nature on any level that has anything like it.

---

another column.

The relationship between L_1 and L_2 is such that the columns one square wide on L_2 correspond to the columns two squares wide on which circles on L_1 are forced to stay. L_1 and L_2 also have the same resolution in terms of moments of time, so that any T_n on L_1 has a directly corresponding T_n on L_2.

Now, for any T_n and T_{n+1} on L_1, every circle on L_2 deterministically moves down one square.

[586]However, I do not define what exactly *would* be a law of nature here, since I am specifically talking about a broader group of rules where it does not matter which of them are laws of nature.

11.6 A complete definition of determinism

Adding mention of level of description L as well as part of the universe P into the definition of determinism gives the most complete definition of determinism, since the previous definitions can be derived from this by assuming L is the ultimate level and P is the entire universe.

> Determinism holds in a part of the universe P at a level of description L between times T_1 and T_2 iff, given any total state of part of the universe P as described in the terms available in L at T_1, and any total input from the rest of the universe between T_1 and T_2, as described in the terms available in L, only one total state of P as described in the terms available in L is possible by the rules of L at T_2.

The important thing to remember is not this exact formulation so much as the general idea that determinism means that only one future state of affairs is nomologically possible, and that this can also apply in a proper part of the universe or on some other level of description than the ultimate one. Having given the definition, I will next make some observations about how this all applies to the free will question.

11.7 Applications to free will

Though the incompatibilist (or compatibilist) position with respect to determinism and free will is usually seen as pertaining to determinism on the ultimate level, a little reflection shows that some higher levels also have to be relevant when speaking of human choices. There is a clear sense in which choices do not happen on the ultimate level at all.

The ultimate level seems to have some relevance because even though we do not actually have a description of the world on the ultimate level, we can always shift to a level of description where we look at what would happen if determinism did (or

does) hold on the ultimate level, or not. This leads to the line of thought where we can see that things could *really* have happened only in one way, and thus we could not have done otherwise in some potentially relevant and even fundamental sense. If some coarse-grained higher-level description shows the choice as indeterministic instead, that potentially has a great deal of relevance, but it does not simply negate the ultimate-level determinism and its potential implications.[587]

Moving upwards in terms of possible levels of description, determinism on some middle level between the ultimate level and the psychological level(s) on which decisions are normally thought of as being made, such as biological or neurological determinism, could indeed be relevant and possibly impede freedom even from a compatibilist point of view. Though I mention this here for the sake of completeness of the current topic, real discussion of it has its own section: 2.6.1 (and see also 9.6.1).

Finally, the kind of level(s) of description where (in)determinism has the most obvious relevance for freedom and decision-making, besides the point about ultimate-level determinism mentioned above, are high-level psychological or intentional levels. Here we talk about motivations, desires, decisions and so on. We cannot, at least in practice, see what happens on the ultimate level when someone is making a decision, and we are far from being able to ever see the correlations between microphysical and as it were decision-level events.[588] When we look at decisions and what led to them, we look at higher-level phenomena. Obviously, then, some kind of higher-level determinism should have relevance for our thoughts about free will.

One way in which high-level determinism is relevant can easily be introduced

[587]Christian List (2019) gives the closest thing I have seen to an argument that higher-level indeterminism really would negate the relevance of determinism on the ultimate level. His view is discussed in this work in sections 1.3 and 3.4.3.

[588]To do this in any strong sense would require traversing an endlessly complex conceptual terrain that has been called "Ant Country" (Stewart & Cohen 1997, chapter 3; see also Kokko 2014, pp. 56–58), so it may be impossible even in principle.

by continuing the thought introduced above. While higher-level indeterminism cannot ensure that we "could have done otherwise" in the sense that seems to worry incompatibilists, not if determinism still holds on the ultimate level, neither can lower-level indeterminism help in the same sense if our decisions are deterministic on the relevant higher level.[589] If our choices are deterministic as choices, then the microphysical indeterminism on the ultimate level is irrelevant to whether we can *choose* otherwise.[590, 591]

What levels are actually relevant, and on which level(s) (in)determinism is required or not for freedom and why, is the topic of much of the discussion of the rest of this work.[592]

11.8 Determinism and exact circumstances

One thing to be understood about definition as it is described, defined and discussed here is that nothing except *exactly* the same initial conditions (including output from the rest of the world) need to produce the same result. If all we know is that determinism holds, small changes in the conditions could produce arbitrarily large results in the results. This could be particularly plausible with respect to human

[589]We have to be careful here, though, since there can be various different high-level levels of description, even ones that are relevant and useful in some sense, that abstract away so much detail that things can appear deterministic in them even though this is not really relevant to whether free will exists or applies. As a very simple example, it is possible to have a level of description with the variable "agent acts freely or not," and if this variable deterministically gains the value "yes," this is tautologically not a threat to freedom. For actual discussion of what levels are relevant and why, see the cross-references in the footnote for the next paragraph.

[590]Mind you, "being able to do/choose otherwise" is a contentious expression (cf. 2.4.2), and in this context we should really speak of it being possible that we choose otherwise. I use this wording for simplicity.

[591]This argument is made a second time with a little more length in section 3.4.4, where it ties in with the main line of argument of chapter 3.

[592]Some of the most relevant sections are 2.6.1, 3.2.2, 3.4.3, 3.4.4, 3.5, 5.4.1, 8.1, 9.1.3, 9.3, and above all 5.4.

action. As a simple example, a mere group of pixels on a screen could make the difference in a whole group of persons getting their things to a car and driving to the train station or not, if the pixels spelled out the message that the train has been cancelled. This example is meant to be seen as a small change on the physical level causing a relatively much larger change on the physical level, but analogously, on a level of description describing decision-making, anything that shows up on that level could in principle lead to any change in the outcome, without threatening determinism on that level. Thus, a deterministically acting person would not need to act the same way in all similar circumstances, merely exactly the same ones.

This is worth mentioning because imagining determinism can easily lead to imagining a more rigid view of what it would mean for behaviour. The idea is applied a number of times elsewhere[593], but for now, it is enough to remember the general point.

11.9 A case study: Thomas Reid as a compatibilist who speaks like a libertarian

As I emphasise repeatedly in 2.1, this definition of determinism and indeterminism is independent of many things that are often associated with the idea of determinism, such as whether causality applies and whether materialism is true. As a concrete example of this, I look in this section at Thomas Reid's philosophy of free will, mainly in his *Essays on the Active Powers of Man*[594], and show how he consistently both affirms a compatibilist position against certain genuinely incompatibilist views, and opposes the kinds of things incompatibilists and libertarians typically want to oppose. Thus, his being classified as a compatibilist by the standards I have set forth above is surprising at first sight, but it makes sense even in terms of whom he

[593]Particularly 2.6, 5.3, and 5.4.

[594]Reid 2010. In the rest of this section, references to this book will be by page or section numbers only, while no other works are referenced, not counting cross-references to the present work.

distinguishes his position from, and it helps set apart what is actually involved in this definition and what is not. Though this may seem a surprising detour in an appendix mostly providing dry definitions, I find this necessary to emphasise due to the amount of confusion that gathers around these points.

Reid's ideas discussed here also foreshadow many of my own ideas, and I note those connections in the footnotes in this section, though I refrain from bogging down the main text in this subsection with metatext pointing them out.

There are three clear reasons why Reid seems – possibly even to himself, though he does not use the term – like a libertarian.

Firstly, Reid speaks against "necessitation", which sounds like it is a form of determinism, presumably the form that was fashionable in his time.[595] Reid's necessitation is definitely something unfree, as we will see below, so it seems implied its proponents are incompatiblists and thus hard determinists.[596] He also opposes Spinoza[597] and Leibniz[598], who are necessitarians by a more modern use of the word that is stronger than determinism, but he is also speaking against simply a mechanistic picture of nature, where causal reasons outside the person cause everything that the person does.[599, 600]

---

[595]It was particularly advocated, particularly in Reid's mind, by Priestley (xx-xxi).

[596]See 231.

[597]252.

[598]212. Reid also targets a form of "the doctrine of necessity" that he associates with both Priestley and Hume's idea that causation is nothing but constant conjunction (212–213).

[599]208.

[600]On the other hand, Reid speaks of "determination", "determining" and "(actions) being determined" as the thing that agents do when choosing, consciously or not (e.g. 64), which if nothing else is somewhat ironic from a modern point of view in the light of his (supposed) opposition to determinism and compatibilism, though this is only a question of word choices, not evidence for his compatibilism.

Secondly, Reid also argues against forms of compatibilism, or at least one form. (He also does not ever say he agrees with any compatibilist.) He disagrees with Locke's and others' idea that freedom is just voluntariness,[601] since that can also mean voluntary acts determined by something other than the person's will in an inappropriate manner[602].

Thirdly, Reid clearly defends some form of **agent causation**, which means that the agent as a "substance" rather than some prior event or even other substance is the cause of the agent's choices.[603] Agent causation is a libertarian idea in modern times, and Reid's characterisation of it seems strongly along these lines as well.

Though this point is dependent on the above three, it is also worth noting that, if told briefly about the current terminology, it seems as though Reid would have identified as a libertarian. He might think that a libertarian position that freedom excludes "being determined by prior factors outside yourself" and "not being able to do otherwise," instead demanding that "the choice is up to the agent," would sound just right.

All of that said, the actual details of Reid's theory are less libertarian, and particularly less incompatibilist, if we look at their relationship with the general definition of determinism explicated in the present work rather than the particular forms he discusses. The case for Reid's incompatibilism and libertarianism is only strong when selectively attending to the most obvious markers, but not when attending to this detail. Next, I will explain in some more detail how Reid thinks choice works, showing these ideas, where they are not simply too vague to say, are compatible with compatibilism.

---

[601] 199-201.

[602] 196; discussed further below.

[603] Agent causation is discussed in the present work as a possible way of avoiding problems of indeterminism in free will in section 3.4.2. Out of Reid's ideas presented here, it is perhaps the ones I am the least sympathetic to, but that is for reasons that are besides the main point about determinism and indeterminism in free will (cf. 2.1.6).

In Reid's metaphysics of causality, persons making decisions are the proper case of causation, and all other talk of causation is only derivative and inaccurate.[604] Reid also says that it is efficient causes that natural scientists have not found in events of nature.[605] So, real causality is left up to human beings who cause their own choices and God who causes everything else.[606] This means Reid believes in agent causation. However, it is possible to at least partly analyse what happens according to Reid's system when an agent causes their own actions, and nothing there says that something is necessarily indeterministic in our sense.

Reid posits "necessitation" as the opposite of "the liberty of a moral agent" as follows:

> If, in any action, he had power to will what he did, or not to will it, in that action he is free. But if, in every voluntary action, the determination of his will be the necessary consequence of something involuntary in the state of his mind, or of something in his external circumstances, he is not free; he has not what I call the liberty of a moral agent, but is subject to necessity.[607]

---

[604]Essay I chapters II, III, V, VI; 203-205; essay IV chapter III; 225, 229-230.

[605]38, essay IV chapter III.

[606]Reid does state that it cannot be known whether human beings are really the effective causes of their actions (or merely the occasional causes), but that they are still clearly the causes in the sense that makes them morally responsible. (41–44.) Later, Reid states that it is more or less inborn to have the intuition that we are the effective causes of our "deliberate and voluntary actions," without confirming that this intuition would be correct but neither pointing out it may not be, thus leaving it sounding at this point as though it is (203–205).

[607]196. 198 affirms that this is the definition he is going to use for "necessity" from then on. Notice how this also explains how he does not accept mere voluntariness as freedom, as noted above in the context of his opposition to Locke. This shows the wording that Reid is opposed to freedom as voluntariness is almost misleading, since the kind of voluntariness he counts as unfree is voluntariness "determined necessarily" by something involuntary, so voluntariness of some sort or lack of it comes up again almost immediately in defining his view as well. Further, this definition shows that Reid's later assertion that the concepts of "necessary cause," "necessary agent" and "acting from necessity" are self-contradictory (212) is not a statement of

Note how this necessitation is defined in terms of something involuntary determining the agent's voluntary actions, not being determined in general. At the same time, it mentions that it must be within the agent's power to will or not will their action. This wording can be opposed to determinism if the ability to do otherwise is analysed as indeterminism, but it is not always interpreted that way. Examining Reid's explanation further, is there anything that indicates the interpretation should be indeterministic?

Reid describes and opposes the necessitarian model in which different "motives" pull a person in different directions, and the strongest motive wins.[608] This could always be made trivially true by saying whatever thing guides a person's choice is the strongest motive.[609] In Reid's view, however, it must be the agent that is causally creating their actions, not a motive, which is not even the right kind of entity to be a cause.[610] Motives are more like advocates making their cases to a judge,

---

incompatibilism.

This definition makes it almost annoyingly easy for him to say things that sound like statements of incompatibilism without understanding of the definition: "If, again, the meaning of the question be, Was there something previous to the action, which made it to be necessarily produced? Every man, who believes that the action was free, will answer to this question in the negative." (246) sounds like a clear incompatibilist statement, but the definition of necessity as something opposed to freedom makes it a mere tautology.

Sometimes, from my limited knowledge, I have to wonder whether Reid is not just using his own definition to rebut arguments that are talking about a different kind of necessity: "But I know no rule of reasoning by which it can be inferred, that, because an event certainly shall be, therefore its production must be necessary." (254.) Of course, his opponents may have been doing the same thing in reverse, jumping from necessity in our sense to unfreedom, smuggling in incompatibilism.

[608]Essay IV chapter IV.

[609]I am not putting this forth entirely as Reid's point here, so much as a truth I am asserting, but he makes about the same point in (216–217).

[610]214, 217.

analogous to the agent, who listens to them but ultimately decides.[611] A motive being strong may simply make it difficult to resist[612], but the person themselves may make the effort to do so. Still, to have meaningful acts of the will, there must be an inclination to do something in the first place[613], even though it is the will rather than the inclination that determines what the person ultimately does.

One indication of how free will works for Reid is given in his statement that, if people indeed had free will, we should expect to see (as we do) that people sometimes deviate from the conduct that is according to best reasons, and those who are more competent at self-control (etc.) do so less often than those who are less competent. This shows that free will is at least roughly statistical on the aggregate level. Given many occasions of people choosing freely, the results will tend towards an average number of right and wrong choices based on each person's competence at choosing.[614]

An agent is free, for Reid, if they can make their decisions as they wish without being determined by causes outside of them. Still, as seen above, more competent agents are more likely to be "determined" by good reasons. A person who is incapable of reasoning at all is not free or responsible. Neither is someone who is compelled by a passion so strong that it is irresistible.[615] Somebody giving in to a

[611]217. In the present work, I discuss this kind of notion as an attempted solution to the problems of libertarianism in section 3.4.5.

[612]56-57, 62, 218, 234–235. See also 137–139.

[613]51, 74-75, 197. Note that this is not the same thing as a motive, which many acts (and even deliberate acts) do not have according to Reid (215–216). I note this point without attempting to clarify what the difference is further.

[614]220. Reid does not speak of specific numbers, but I take his talk of greater and lesser numbers to imply something like this. Nothing here really hangs on whether the statistical interpretation is exactly correct or should be replaced by something more vague but still numbers-based. See also 199.

[615]52–53, 56–57, 59, 142, 224–225.

temptation to do wrong is responsible, though if the temptation was strong and hard to resist, they are proportionately less responsible[616]. Also, someone who does the right thing out of mere inclination and not by reasoning and following the correct principles is not praiseworthy,[617] and they cannot be *trusted* to continue to do the right thing, since they do not have principles ensuring this.[618]

Meanwhile, someone who was "determined" by the right reasons through their reasoning and agent causality is free – even if that actually implies determinism, as is seen in the case of God.

Whereas elsewhere, Reid's theory merely does not contradict compatibilism – everything that he calls freedom could happen in a deterministic system – with God's freedom he quite explicitly says that God, as a perfect decision-maker, would unfailingly do the right thing based on his reasons. This does not match the mechanistic or physicalistic stereotype of determinism, such as Reid explicitly opposes, but it matches the simple definition of determinism: prior conditions and the way things work make it so that only one way God chooses is possible.

Others have considered it a problem that God cannot be absolutely free if he is only able to do what is best because he is morally perfect, and thus cannot do otherwise. Reid, on the other hand, approaches the matter like a compatibilist objecting to the principle of alternative possibilities. God has a perfect competency to act morally, and so God will act always morally. According to Reid, to say this is unfree would be to say that the very thing that makes God perfect makes him unfree.

---

[616] 57, 59.

[617] 165–166.

[618] 67–69. Cf. 94–95, 102, 164–165. I suppose that this can be said this way in the modern sense of "principle", like a rule, even though Reid himself tends to use the word "principle" to mean something like any cause from which actions originate (see especially essay III part I chapter I). Compare this with my response to "The Lady, or the Tiger?" in 3.4.7.

> Rational beings, in proportion as they are wise and good, will act according to the best motives; and every rational being, who does otherwise, abuses his liberty. The most perfect being, in every thing where there is a right and a wrong, a better and a worse, always infallibly acts according to the best motives. This indeed is little else than an identical proposition: For it is a contradiction to say, That a perfect being does what is wrong or unreasonable. But to say, that he does not act freely, because he always does what is best, is to say, That the proper use of liberty destroys liberty, and that liberty consists only in its abuse.[619]

Not only is this *like* compatibilism, it outright *is* compatibilism. God's acting freely from the best reasons is not determination by strongest motive. However, it is not only compatible with the idea, but it demands that God's best actions are determined unambiguously by what came before. In this case, what came before is described in terms of reasons, but there is no obstacle to describing it in parallel in terms of the state of the world.

Reid's idea of agent causality would presumably prevent describing it in terms of antecedent *causes*, with only God being recognised as the cause, but this is just a distraction since the central idea of determinism is not dependent on causation as such, and incompatibilists would by definition not be happy to be compatibilist about non-causal determinism. It might also be arguable that agent causality with no causal determination by what came before but nevertheless determined by antecedent conditions could equally well be analysed in terms of deterministic causality – in other words, that to deny it being causality is not even a meaningful position – but I will not spend effort defending that idea when I do not need it to come to my conclusion.

Making matters even clearer, Reid says quite explicitly that God's actions would be in principle absolutely predictable in this way, and he does this while arguing against the notion that free actions in general cannot be predictable:

---

[619] 214–215.

> With regard to the general proposition, That it is impossible that any free action can be certainly foreseen, I observe,
>
> *First*, That every man who believes the Deity to be a free agent, must believe that this proposition [that free actions cannot be predictable] not only is incapable of proof, but that it is certainly false: For the man himself foresees, that the Judge of all the earth will always do what is right, and that he will fulfil whatever he has promised; and, at the same time, believes, that, in doing what is right, and in fulfilling his promises, the Deity acts with the most perfect freedom.[620]

The difference between God and humans in this respect is nothing such that humans would be unfree if they were similarly predictable. As mentioned above, more competent deciders more reliably follow the objective reasons to do what is right.[621] God is merely the perfect version of this, the same as the perfect human would be,[622] and this is freedom, not lack of it. Nowhere does Reid say that being determined in *this* way is unfree. On the contrary, he calls it perfect freedom, as seen above. Yet, in the basic sense, it is determinism.

Thus, Reid's ideal free agent is deterministic by definition. Those who are less free are less determined by their reason, and those who are not free at all are

[620]255. Italics in original.

[621]See also 68: "We may observe, that men who have exercised their rational powers, are generally governed in their opinions by fixed principles of belief; and men who have made the greatest advance in self-government, are governed, in their practice, by general fixed purposes. Without the former, there would be no steadiness and consistence in our belief; nor without the latter, in our conduct." Compare also how, in essay IV chapter VIII, Reid argues that the ability of humans to rationally follow plans is proof of their freedom; clearly this is freedom as rationality, not as the possibility of doing otherwise regardless.

[622]"He only who does in all cases what he ought to do, is the perfect man." (192.)

determined by something else without their reason having any chance to interfere. It is not a question of freedom being incompatible with determinism, but freedom being associated with one kind of determinism, that based on the right reasons – and *other* kinds of determinism contradict *that*.

Thus, Reid explicitly opposes all of the following: "necessitation" in a mechanistic hard determinist sense, free choices being determined as part of a chain of event causation or by something outside the agent; mere voluntarism; and lack of "contingency" in some sense in (free) choices. And in spite of this, he is in favour of choices being the more perfect the more they are determined by the agent's good reasons, and he explicitly says it is nonsense to suppose God's freedom would be compromised by his perfection, while the perfection makes God, frankly, deterministic.

Shortly put, Thomas Reid is motivated by many of the same motives that seem to motivate libertarians today, and he speaks in similar if now old-fashioned terms. Yet he is still compatibilist by my simple definition, because he does not think free choices have the property that they might always happen otherwise, as that would only lead to irrationality and imperfection. That is the conclusion that follows from the definition of determinism that I am actually using. Determinism is when only one thing can possibly happen next – nothing else, less, or more.

12 Appendix C: An overview of the randomness argument

This appendix summarises the claims and sub-claims of the randomness argument as it is made in chapter 3, with references to where they are explained in more detail in the chapter.

- **Main claim 1: Indeterminism contradicts concrete (practical) control, and concrete control is very important for independent reasons.**
- **Main claim 2: There is no form of indeterministic freedom that is independently desirable – that is, aside from desires aimed at having indeterminism itself.**
- This argument works specifically with the definition of determinism and indeterminism that is given in this work (2.1 and Appendix B).
    - The argument cannot be bypassed by talking about kinds of causality or the presence or absence of causality. Vary a situation in terms of kinds of causality but keep it the same in terms of (in)determinism, and you will have achieved nothing in terms of the randomness argument. (3.4.2.)
- Main claim 1 concerns concrete control specifically; however, when evaluating main claim 2, other claimed forms of control and other consequences of indeterminism can be evaluated for desirability. (3.5.)
    - Arguing that indeterminism allows agency in some sense that does not secure concrete control does not counter main claim 1. (3.4.1; 3.4.2.)
    - There are likely cases where randomness is concretely useful, but there are none where indeterminism on the ultimate level is required rather than

pseudorandomness, and hence, this is no evidence for the independent desirability of indeterminism. (3.5.)

- There is no third option besides determinism and indeterminism. (3.2.1; 3.3.7.)
  - Even if a system is a black box that we cannot see inside, its output is either deterministic or indeterministic.
  - It is not possible to appeal to an agent's choice at any part of the process without explaining whether it is deterministic or indeterministic.
- The randomness argument does not need complete randomness to work. If indeterminism plays a small role at some part of the process, then that specific point in the process is random in the relevant sense. (3.3.5.)
  - The argument applies both when indeterminism plays a small role in terms of only appearing in one part of the process, and when it plays a small role in terms of being only partial (e.g. statistical) when it does appear.
  - Indeterminism may play such a small role in choice that, in spite of the randomness argument, it does no harm; however, there is no place where it can do anything that determinism cannot that would make it a better option for any other reason than it being postulated that indeterminism is needed.
- Indeterminism might only apply to some choices; in that case, it either causes problems in those cases, or it is constrained only to choices that are unimportant in the sense that which thing is chosen is unimportant. (3.3.6.)
- Ultimate origination as defined by incompatibilists ends up being equivalent for current purposes to some form of indeterminism, so all the same arguments apply to it. (3.6.)
  - Indeterministic ultimate origination would mean loss of concrete control.
  - Full ultimate origination is self-contradictory, so it cannot be had under indeterminism any better than under determinism.
  - Desiring indeterministic ultimate origination amounts to desiring indeterminism, thus does not give a separate reason to desire indeterminism.

13 Appendix D: Glossary of Terms

It is important, in philosophy even more than elsewhere, to know what one is talking about when using a particular word. Aside from the most central terms *free will* and *responsibility*, I endeavour in this work to use key terms always in the same sense and to make that sense explicit. In this glossary, I give those definitions briefly. Naturally, when I give a very specific definition to a term like "determinism", I am not claiming that this is the correct or only definition for the term overall. There may be other definitions, but these are the ones that I am using when I use a word here without further explanation.

Obviously, I cannot define the whole language that I use. Terms I do not give a definition here may not be used with such precision, but rather as it is normal in language (though I still avoid speaking as if I am speaking about a thing when I speak of the possible meanings of a word; see 1.5, p). That said, this glossary also includes a few terms that I use *without* the intention of using them in a specific sense; I include them to explain that I do not. These are marked with an asterisk *.

Agency perspective: A way of looking at the world where we pay attention to the presence of agents and make moral judgements. Contrasts with *mechanistic perspective*.

Capacity responsibility: Having the capacity to do some thing such that one can be held responsible for doing it in terms of *liability responsibility*.

Causal responsibility: The quality of being the cause of something. For agents, this

also involves having been the cause via an act, not merely that one's body as an object happened to cause something if one was not in control of what one's body did.

The challenge of freedom: Our psychology is such that we are unfree in many ways under the definition of *free will* that I end up choosing, and we can never be completely free. Still, it is possible for us to become more free by developing our skills. Thus, the challenge is always to become more free.

The challenge of responsibility: Holding someone responsible may be empowering or otherwise lead to good results if it is done at the right time and the right way, but it may be demoralising and have other bad consequences if done wrong. Conversely, not holding someone responsible may be liberating, or alternatively lead to irresponsibility, disempowerment or both. The challenge of responsibility is knowing when and how apply or not apply it such that good rather than bad things follow.

Compatibilism: The stance that free will is compatible with *determinism*.

The compatibility question: The question of whether free will is compatible with *determinism* or not.

Concrete control: (Also practical control.) A kind of control an agent can have over their own actions where the agent can rely on the actions being derived from the agent's own reasons and deliberative processes, on the agent's future actions being predictable by the agent to some degree, and on being able to act more or less rationally. It can be shown that (local) *determinism* is a necessary but not sufficient condition for concrete control, and hence that *libertarian* theories of free will are broadly incompatible with it, leading to the *randomness argument*. Concrete control also cannot be reconciled with *ultimate origination*.

Determinism: The possible state of affairs that the universe or some part of it is

governed by laws that determine exactly what is going to happen next, with no alternative possibilities. Unless otherwise specific, refers to universal determinism. If specified specifically as affecting a part of the universe, means local determinism. In both cases, unless otherwise specified, refers to determinism at the *ultimate level of description*.

Appendix B is dedicated to explaining how this concept is defined in this work, so I only reproduce the most comprehensive definition here and refer the reader to said appendix for explanation of the details.

> Determinism holds in a part of the universe P at a level of description L between times T_1 and T_2 iff, given any total state of part of the universe P as described in the terms available in L at T_1, and any total input from the rest of the universe between T_1 and T_2, as described in the terms available in L, only one total state of P as described in the terms available in L is possible by the rules of L at T_2.

Emergence: Unless otherwise specified, refers to *weak emergence*. See also *strong emergence*.

***Free will:** Because this is one of the key terms discussed in this work, and the focus of this work is on to understand how these key terms should be understood, "free will" is *not* used consistently within this work. Thus, I cannot give it a definition here that the reader could apply in different parts of the text to interpret the term, especially where the text discusses the ideas of others. The following is the definition I arrive at at the conclusion of my arguments: An agent *A* has free will to the extent that, when making choices, *A* decides in accordance with *A's* interests all things considered.

Full ultimate origination: See *ultimate origination*.

Fundamental level: The same as *ultimate level* (unless used relatively, i.e. "more fundamental").

General reasons-responsiveness: A quality that an agent has to the extent it can be said that their choices are responsive to good reasons they have. Thus, it is a sliding (and necessarily vague) scale.

Hard determinism: The stance that free will does not exist because free will is incompatible with *determinism* and the world is deterministic.

Hard incompatibilism: The stance that free will is impossible because it is incompatible with both *determinism* and *indeterminism*.

Incompatibilism: The stance that free will is incompatible with *determinism*.

Indeterminism: The state of affairs in which *determinism* does not hold. All further specifications of kinds of indeterminism mean the denial of determinism with the same specifications (e.g. local indeterminism in regards to a part of the universe means the denial of local determinism in that part – though from it also follows the denial of universal determinism in that universe). See *determinism*.

The intelligibility question: The question of whether free will is compatible with *indeterminism* or not.

Intrinsic-good retributivism: The idea that punishing (appropriately harming) people who have done something wrong is, in the right circumstances, morally good in itself. I also use this term to cover cases where this is what is assumed in practice, even if it is not outright stated in those words or it is even verbally denied.

Intuition: ***1.*** Any belief or inclination to believe in something that comes to a

person's consciousness directly, not based on explicit reasons. ***2.*** Whatever collection of processes or properties causes the person to have intuitions in sense 1.

Leeway incompatibilism: *Incompatibilism* based on the idea that *determinism* does not allow multiple opportunities when making a choice.

Level of description: A way of describing (a part of) the world that involves particular kinds of entities and possibly rules for how their situation changes with time. See Appendix A for a thorough description.

Liability responsibility: Being held liable, morally or otherwise, to face consequences based on what one has done.

Libertarianism: The belief that free will exists and is incompatible with *determinism*.

Local determinism: *Determinism* as applying only to some proper part of the universe.

Mechanistic perspective: A way of looking at the world that only considers events and not agency. Contrasts with *agency perspective.*

The mode of agency: A way of looking at events in the world that people seem to gravitate towards, in which agents take actions and are responsible for them, and explanations about the causes of the actions are not traced beyond the agent's decision. Contrasts with *the mode of randomness* and *the mode of causality*.

The mode of causality: A way of looking at events in the world that people seem to gravitate towards, in which things happen for simple causal reasons and nobody has agency. Contrasts with *the mode of agency* and *the mode of randomness*.

The mode of randomness: A way of looking at events in the world that people seem to gravitate towards, in which things happen for no reason or no known reason and nobody has agency. Contrasts with *the mode of agency* and *the mode of causality*.

***Moral responsibility:** Responsibility of various sorts to the extent that it is judged morally. See especially *role responsibility*. What exactly this means is not defined from the start because how to understand it is one of the main questions of this work. The general normative definition of (moral) responsibility arrived at in this work is: An agent *A* is morally responsible for morally relevant state of affairs *S* (past, present, or future) to the extent that *A* had/has/will have a reasonable chance of affecting whether *S* occurs or not. *A* is morally responsible for an act *B* to the extent that *A* has a reasonable opportunity to do or not do *B* and that *B* affects whether *S* occurs.

The myth of pure evil: The implicit assumption, fed by subtle psychological bias, that evil (harmful) acts are only committed by people who are especially malicious of character.

Outcome responsibility: Refers to when an event is held as attributable to an agent's actions. It requires both causal responsibility – the agent must have played a causal role in the event happening, of course – and role responsibility.

Postulated ownership: When an agent has postulated ownership of an action or choice, that action or choice is considered as the agent's own action or choice, and originating with the agent, regardless of any other factors about it. The ownership or origination can be seen as meaning something that has implications for freedom and/or responsibility, but as inherent properties of the action itself, they cannot be explained non-circularly due to the nature of this concept (the "regardless of any other factors" clause). Contrast with *concrete control*.

Practical control: Another name for *concrete control*.

Pseudorandomness: *Indeterminism* on a higher *level of description* in spite of *determinism* on the lower level or *ultimate level*.

***Randomness:** Not used in the present work as a term with a specific definition but as a word characterising the consequences of *indeterminism*. *Pseudorandomness* also has the same kind of consequences. See section 3.3.3. See also *randomness argument*.

The randomness argument: See Appendix C for a summary.

Reasons-responsiveness: The quality of an agent being responsive to reasons they have in making choices. When used without qualification in this work, implies *general reasons-responsiveness* and *sufficient reasons-responsiveness*.

Retributivism: The idea that it is morally right to harm (punish) people who have done something bad and have *liability responsibility* for that deed. Generally, as seen here, collapses into *intrinsic-good retributivism*.

Role responsibility: What someone is held responsible (morally or otherwise) to do or not do based on who their circumstances – including actually roles, like being a ship's captain and thus having certain responsibilities, but also very general notions like being responsible not to attack others without provocation, and *ad hoc* assignments of responsibility such as a particular person being held responsible to do a particular thing on a particular occasion

Scrutability: A class of truths *B* is scrutable from a class of truths *A* iff, if an agent knew all truths of type *A*, then the agent could know all truths of class *B* based on that.

Smilansky's ultimate level: A way of describing the world, supposed to be true and relevant by Saul Smilansky[623], that reveals a deeper truth than any other and on which there exists no full freedom or responsibility, only causal events. I dispute that this is a correct way of looking at things; see *mechanistic perspective*. Not to be confused with my *ultimate level of description*.

Source incompatibilism: Incompatibilism based on the idea that determinism implies that one's choice are caused in the wrong way, not by the agent themselves.

Strong emergence: If *level of description* L_2 is strongly emergent from level L_2, there is a dependence between the levels, and L_2 cannot exist without L_1, but L_2 also contains something that is not *scrutable* from L_1 even in principle (without extra knowledge about how the levels relate to each other that would not be needed to describe L_1 alone).

Sufficient reasons-responsiveness: The quality of an agent being responsive enough, in making choices, to good reasons they have, that it makes sense to consider them as responsible in the relevant context.

Ultimate origination: An agent has ultimate control over their actions or choices if they are the ultimate origin of those actions or choices, meaning there is nothing prior that determines them. In this work, I concede (for convenience) to call it "ultimate origination" when choices are undetermined and under the agent's control in some metaphysical sense such as agent causation. *Full ultimate origination* where the agent's choices are not caused by anything prior and are controlled by the agent in a more robust sense, including concrete control, is self-contradictory because

[623]Smilansky 2000.

indeterminism contradicts *concrete control*.

Ultimate level (of description): A hypothetical *level of description* that describes the universe in perfect (physical?) detail such that every other true level of description is in principle derivable (or scrutable; see *scrutability*) from it. The description of laws of nature applied at this level, assuming it can exist within physics, is known as the Theory of Everything. If there is more than one equivalent level of description fulfilling these criteria, they are all treated as one here nevertheless, since they are equivalent. Not to be confused with *Smilansky's ultimate level*.

Universal determinism: See when *determinism* applies to the whole universe.

Weak emergence: If *level of description* L_2 is weakly emergent from level L_2, there is a dependence between the levels, L_2 cannot exist without L_1, and L_2 is *scrutable* from L_1 in principle (without extra knowledge about how the levels relate to each other that would not be needed to describe L_1 alone).

It is common for the question of free will to be posed as "Do we have free will, or is everything determined in advance?" However, this is not the question that is usually posed in philosophy. Rather, the philosophical debate is often about whether or not free will is compatible with determinism: whether it can be true at the same time that we have free will and that everything that happens is determined by that which comes before it. One reason is that on a closer examination, indeterminism also seems problematic: if nothing fully determines your choice, it seems you are not fully in control of it either. The same questions also seem to apply to the concept of moral responsibility: it seems we cannot be responsible if we either "cannot do otherwise" or "are not in control." Besides, free will itself is usually seen as a precondition for moral responsibility.

This dissertation asks the question of how we should best understand the concepts of free will and moral responsibility and their relationship to each other. After a careful analysis of free will and (in)determinism, it turns out that, despite some intuitions to the contrary, determinism in general poses no concrete threat to free will, but indeterminism does. However, the wrong kind of determinism is indeed a threat, and this reveals more about how free will should be best understood.

From this, the study continues to moral responsibility. Many things are often taken for granted about moral responsibility, particularly its relationship with free will and its justifying punishment. However, these claims are much too important to be accepted based only on intuition. Instead, they can be justified by ethical arguments starting from the kind of view of free will that is developed in this work. This leads to a unified theory of how we should understand free will and responsibility together.

Acta Philosophica Turkuensia

University of Turku
COIMBRA GROUP

ISSN 3087-5943
ISBN 978-952-02-0440-2 (PRINT)
ISBN 978-952-02-0441-9 (PDF)
Painosalama, Turku, Finland 2025