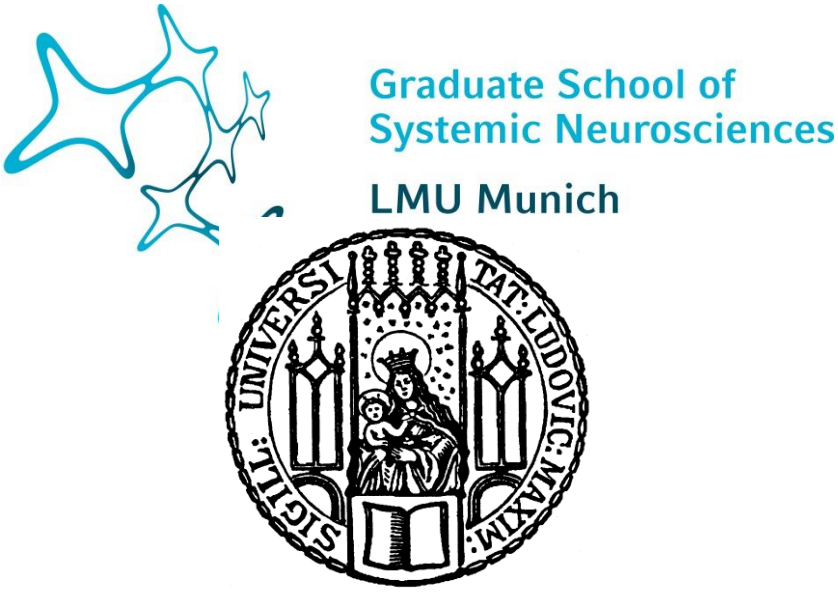
Freedom and Chance: Towards a Graded Approach to Free Will

Mark Wulff Carstensen



Dissertation at the
Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München

May, 2021

Supervisors
Prof. Dr. Ophelia Deroy
Chair of Philosophy of Mind
Ludwig-Maximillian-Universität München

Prof. Dr. Stephan Sellmaier
Research Center for Neurophilosophy and Ethics of Neuroscience
Ludwig-Maximillian-Universität München

Prof. Dr. Paul C. J. Taylor
Munich Center for Neuroscience
Ludwig-Maximillian-Universität München

First Reviewer: Prof. Dr. Ophelia Deroy
Second Reviewer: Prof. Stephan sellmaier
External Reviewer: Prof. Gregg Caruso

Date of Submission: 04.05.2021
Date of Defense: 12.08.2021

Acknowledgements

First and foremost I want to thank Ophelia Deroy for her support, guidance, feedback, and kindness as a supervisor. She has been a big part of what made the whole ordeal of doing a Ph.D. manageable and, at times, even enjoyable. Her everlasting efforts to instill good habits in my writing, treatment of data, and general approach to the subject matter have undoubtedly contributed crucially to the quality of my work. Secondly, I want to thank Stephan Sellmaier for his encouraging attitude, and for always having time for feedback and discussions. Many thanks to Paul Taylor as well for his feedback and input. He has been a great help in expanding my own views and in navigating the scientific landscape.

I would like to thank the members of the Cognition Values and Behaviour group. It is a wonderfully constructive interdisciplinary group. I would have learned less than half as much if it had not been for the continuous feedback and assistance of the CVBE members. I also extend my gratitude to the members of the Research Center for Neurophilosophy and Ethics of Neuroscience at LMU Munich for their engaged notes and comments throughout the years. Finally, I would like to thank the Graduate School of Systemic Neuroscience for their financial and social support, and for providing the opportunity to partake in a stimulating and truly interdisciplinary work environment. A special thanks go out to the GSN administration – especially Lena, Stefanie, Renate, and Nadine - for always being of assistance and for making the process go as smoothly as possible at every stage.

I would also like to thank Gregg Caruso for being my external reviewer and for being part of my examination committee. I was happy when I learned he would be reviewing my thesis as I admire his work which is a core part of the movement that motivated the approach of this thesis in the first place.
I also thank Markus Paulus for being part of my examination committee.

Summary

The problem of free will is a long-standing and controversial problem in philosophy. Though there are many angles to the question of whether human beings have free will or not, one of the key contemporary challenges is 'the problem of luck'. The problem challenges the idea that we can have free will in an ontologically indeterministic world, by raising the objection that ontological indeterminism in the mechanisms of human decision-making introduces an element of freedom-undermining randomness. In light of contemporary probabilistic models from science, the problem of luck is becoming increasingly more relevant. This problem serves as the core motivation of this thesis.

The primary research question asks how probabilistic uncertainties in the mechanisms underlying intentional action should affect our understanding of free will. The thesis consists of three parts, which in combination aim to address this overarching question.

The first part asks first what the ordinary concept of free will is, and secondly whether the common-sense notion of free will changes with differing beliefs about whether substance dualism or physicalism is true. I.e. whether different beliefs about physicalism drive different free will conceptions. These two questions are explored through two empirical experiments. The first experiment investigates whether the folk-concept of free will is compatibilist, incompatibilist, or hard incompatibilist. Our findings support the hypothesis that the common-sense notion of free will is incompatibilist, but not hard incompatibilist. That is, laypeople conceive free will as incompatible with determinism, but compatible with the presence of probabilistic chance. The second experiment examines if more physicalist beliefs drive more incompatibilist and hard incompatibilist conceptions of free will. While we find support for *apparent* differences in free will conceptions based on beliefs in physicalism, this difference is shown to be driven by different understandings of determinism: Physicalist beliefs drive epiphenomenal interpretations of determinism. Such interpretations are at odds with the philosophical concept of determinism. When this understanding error is corrected, we find no support for genuine differences in free will conceptions between dualists and physicalists. Whether people believe in physicalism or substance-dualism, the common-sense notion of free will remains incompatibilist but not hard incompatibilist.

The second part asks if we have good reasons to accept ontological indeterminism. Three different epistemological approaches to the question are considered: Direct scientific realism, virtue epistemology, and epistemic pragmatism. It is argued that regardless of which epistemological approach we choose, we have sufficient reason to accept ontological indeterminism. Particular emphasis is put on

the approach from epistemic virtues. It is argued that to satisfy the virtue of being genuinely self-critical, we must accept ontological indeterminism in light of best-explanation probabilistic models.

The third part asks what kind of control indeterministic agents can have that would alleviate the problem of luck. A reconceptualization of free will is proposed, which builds upon the common-sense notion of free will. The problem of luck, it is argued, confronts us with a conceptual choice: Either we must abandon the idea of free will for ontologically indeterministic agents or we must accept an element of randomness as part of freedom. Choosing the second horn, it is proposed that indeterministic agents with the capacity for *recursive decision-making*, i.e. decisions that target and alter the agent's own motivational structure, have an important kind of control over their own internal randomness. Although luck remains to some degree, it is argued that having the capacity for recursive decisions nevertheless reduces the problem. As a consequence of this reconceptualization, it is argued that free will and agency is best understood as a graded phenomenon.

Table of Contents

1. General Introduction

"...as absence of necessity is characteristic of what is free, the latter would have to be dependent on absolutely no cause at all, and consequently would have to be defined as *absolutely contingent*. This is an extremely problematic concept, one whose conceivability I cannot vouch for, and one which nevertheless coincides in a curious way with the concept of *freedom*." – Arthur Schopenhauer (Schopenhauer 1839)

Are we in control of how we act? What if everything we do is the outcome of probabilistic, chancy, processes? It certainly feels like we are in control of how we act at least some of the time, and that what we do and what we *choose* to do is 'up to us'. This idea is a core part of our self-conception as persons, and many of our social practices such as law and the whole endeavour of ethics and moral responsibility are often thought to depend upon it. And yet, it is by no means a trivial matter whether our choices and actions are *really* 'up to us', and what exactly this entails. The question of whether we are ever really in control of how we act, and what 'being in control' involves, constitutes the core of the free will problem.

Key definitions:

Determinism: At any instant, exactly one possible future is compatible with the laws of nature and the state of the world at that instant.

Indeterminism: The negation of determinism. I.e. at any instant, more than one possible future is compatible with the laws of nature and the state of the world at that instant.

Compatibilism: The idea that free will is compatible with determinism.

Incompatibilism: The idea that free will is incompatible with determinism.

Hard Incompatibilism: The idea that free will is incompatible with both determinism and indeterminism.

Traditionally, the focus of the free will problem has mostly been on whether free will is compatible with determinism or not. However, in the last century or so, scientific discoveries have increasingly thrown doubt on the traditional view of the natural world as a deterministic machinery. First, with the discovery of quantum physics at the fundamental level of subatomic particles at the turn of the last century, and more recently with the discoveries of contemporary neuroscience which commonly model the biological processes of the brain as probabilistic, indeterministic, mechanisms (see section 1.2 and chapter 4, section 4.1). As stochastic models find increasing success in science, and especially in the sciences related to the mechanisms of decision-making and volitional actions, the problem of free will and determinism grow less relevant in our contemporary landscape. Instead, another angle of the free will

problem gains increasing relevance: The question of whether free will is compatible with *indeterminism*. If everything we choose and do is the consequence of probabilistic happenings, is it possible for us to have free will? Sceptics argue that to introduce undetermined events into the decision-making process introduces a freedom-undermining element of chance and randomness. This objection, known as the problem of luck, maintains that undetermined actions are not under the agent's control and therefore if indeterminism is true, the agent's actions are not under their control and so there is no free will. The work of this thesis is motivated by this consideration. Specifically, I here investigate what the common-sense notion of free will is, I examine if we have good reasons to accept the natural world as ontologically indeterministic – including operations at the level of the brain, and I provide an account of how indeterministic agents can have control of their own internal chanciness. My aim is to contribute to discussions of free will at the intersection between philosophy, psychology, and cognitive neuroscience, by providing an account of free will that is sensitive to the contemporary scientific landscape and the ordinary understanding of free will, and which does not neglect the problem of luck.

Though the problem of free will is intrinsically tied to questions of responsibility, accountability, and morality, one of the main limitations of this thesis is that it will not directly explore the ethical implications of the account presented here. Instead, the focus is to establish what counts as free will ordinarily understood and investigate what kind of control indeterministic agents can have in light of the problem of luck. Whether it leaves room for moral responsibility – and what kind of responsibility it leaves room for – as well as the further social and practical implications of the account are left as an open question for future endeavours (see conclusion). While some thinkers argue that an absence of free will means losing responsibility and attempt to find justifications for free will (Smilansky 2000, List 2019), others argue that moral responsibility can be detached from questions about free will (Frankfurt 1971), while still others argue that giving up on certain ideas of free will and moral responsibility can improve our views on accountability and lead to more constructive and humane social practices (Pereboom 2001, 2013, Caruso 2016, 2019a, 2019b). Here, the indirect contribution is to stress that there is room for free will as it is ordinarily understood within a contemporary neuroscientific view of the world.

The overarching research question asks how probabilistic uncertainties in the mechanisms underlying intentional action should affect our understanding of free will. Each of the four papers that make up the chapters of this thesis includes introductions of their own. The remainder of this general introduction will therefore be limited to providing the general metaphysical and methodological background for what's to come.

1.1 Metaphysical framework

One metaphysical aspect that almost all accounts of free will agree upon is that free will requires mental causation. I.e. if the mind and mental states have no causal effects then free will is impossible[1]. In other words, if epiphenomenalism[2] is true then free will is false. Usually, it is the particulars about what more is or isn't required for free will, which sparks disagreements among philosophers. These disagreements concern mostly differences related to three different aspects of metaphysics that are commonly discussed in conjunction with each other. 1) The question of substance: Whether reality consists of two substances, the mental and material (substance-dualism), or if reality is fundamentally one substance (substance-monism). 2) The question of causal chains: Whether the processes of human decision-making and subsequent actions follow with *necessity* from the history of the universe (determinism)[3], or if these processes include an element of innate and spontaneous chance (Indeterminism)[4]. 3) The question of mechanistic agency: Whether agentive action is reducible to the occurrence and interaction of certain kinds of events – usually agent-involving mental states and events - (event-causal accounts) or if agentive action entails an irreducible agentive component causing the action (agent-causal accounts).

Every position on each of these metaphysical questions comes with their own problems and challenges, though they are often bundled together and discussed in an intermingled fashion. On the question of substance, substance dualism is most famously confronted with the problem of interaction (Elisabeth of Bohemia, see Shapiro 1999, Ryle 1949) but has more recently also been met with empirical challenges (Churchland 1984). Substance monism is confronted with the threat of epiphenomenalism from causal overdetermination and the challenge of non-reductive supervenience (Churchland 1986, Kim 1992, 2005, Tse 2011, List 2019). On the question of causal chains, determinism is confronted with the consequence argument (Ginet 1966, 1990, Van Inwagen 1975, 1983, Fischer 1994) and indeterminism is confronted with the problem of luck (Strawson 1986, Kane 1999, Mele 1999, 2006, Pereboom 2001, Almeida & Bernstein 2003, Sommers 2009, Caruso 2012, 2019b). On the question of mechanistic agency, mechanistic accounts are confronted with the problem of causal deviance

[1] But for an example of free will that does not require consciousness see Brembs 2011.
[2] Epiphenomenalism is the idea that mental states play no causal role in the world (Robinson 2019).
[3] I.e. that only one possible future is compatible with the laws of nature and the state of the world at any given instant.
[4] I.e. that more than one future is compatible with the laws of nature and the state of the universe at any given instant.

(Chisholm 1966, Davidson 1973, Mele 1992a, Schlosser 2007) and the problem of the disappearing agent[5] (Melden 1961, Nagel 1986, Velleman 1992, Lowe 2008, Steward 2013), whereas irreducibly agentive accounts are confronted with the homunculus problem (Kenny 1971), the problem of substance-causality, the problem of non-explanation (Broad 1952, Taylor 1966, Davidson 1971, Van Inwagen 1983, Nagel 1986, Clarke 2003, Schlosser 2010, Pereboom 2014b) and empirical challenges (Chisholm 1964, Bok 1998, Ginet 1990, 1997, Clarke 1996, Pereboom 2001, 2014a).

Though an exhaustive account of free will should elaborate on each of these aspects and tackle the challenges relevant for its respective positions, tackling all the challenges for a complete account of free will is too wide a scope even for a doctoral thesis. Certain limiting choices must be made about the surrounding metaphysical framework that serves as the background for the focus of the thesis. For the purpose of this thesis, I will therefore assume a substance-monist event-causal framework. The focus will be, as noted earlier, the problem of luck and as such the position of indeterminism regarding the question of causal chains.

'Substance monism' is the view that everything consists of fundamentally one substance. Although there are different variants of substance monism, such as neutral monism (Spinoza 1677), idealism (Berkeley 1710), and physicalism (Neurath 1931, Carnap 1959), most contemporary philosophers are physicalists (Bourget & Chalmers 2013), which is the view that everything is fundamentally physical. I.e. that everything fundamentally consists of- or supervenes upon physical 'stuff'. Although there is a controversial and interesting question about whether this position leaves room for the mental and what role mental states and events could play in such a framework, the view I assume for the purpose of this thesis is a non-reductive physicalist view. The main idea is that everything is fundamentally physical but that consciousness and mental states exist as ontologically real and distinct *properties* of the physical (substance monism with property dualism). 'Event-causal' means simply that I accept explanations of agency in terms of agent-involving states and events as full and exhaustive, without reference to any irreducible agent-substance doing the causing.

[5] It is common for the disappearing agent objection to be lumped together with the problem of luck, or even be presented as a version of the luck problem (see O'connor 2007, Pereboom 2014a, 2017). Nevertheless, I prefer to keep the two problems distinct as the disappearing agent objection can be more generally raised against any event-causal account (Mele 2003, 2015), whereas the problem of luck relates only to indeterministic accounts.

1.2 Neuroscience, variability, and probabilities

Neuroscience is ripe with probabilities (Manwani & Koch 1999, Faisal et al. 2008, Ermentrout et al. 2008). At every level of neuroscientific explanations probabilities are employed. Consider the biological mechanisms involved in generating action potentials. Here is a simplified description of how it happens:

Neurons transmit information along their axon by generating action potentials. The neuron's membrane is embedded with ion channels: Small proteins whose gates can open or close to allow ions to flow into and out of the cell. When the membrane potential[6] is at or near the resting potential[7] of the cell, such gates are mostly closed[8]. When the gates open, ions can flow into or out of the cell. If positively charged ions flow into the cell or negatively charged ions flow out, this changes the membrane potential to less negative or even positive values and is known as "depolarization". The opposite of depolarization, i.e. when the membrane potential changes to even more negative values than the resting potential, is known as "hyperpolarization". When the neuron is depolarized enough to reach a threshold level (threshold potential), a positive feedback process is triggered. Voltage-gated sodium channels open which further increases the influx of sodium ions, which depolarizes the cell causing more sodium channels to open and so on. When this positive feedback process is triggered an action potential is generated (the neuron 'fires' or 'spikes'). The action potential traverses along the axon to the axon terminal located at the end of the axon. Immediately after the neuron has fired, the membrane potential drops. The neuron becomes hyperpolarized and enters a refractory period where it is impossible (absolute refractory period) or very difficult (relative refractory period) for the neuron to fire again. Shortly thereafter it returns to its resting state and is ready to fire again. At the axon terminal, when the action potential arrives, ion channels in the presynaptic membrane open, and calcium enters the cell. This causes synaptic vesicles, which are small containers carrying neurotransmitters, to release their neurotransmitters into the synaptic cleft (the gap between the presynaptic and the postsynaptic cell). The neurotransmitters disperse in the synaptic cleft and bind to receptors, such as ion channels, at the postsynaptic cell. When a neurotransmitter binds to a gate, the gate opens, and ions can flow into the postsynaptic cell thereby depolarizing it (or hyperpolarizing it, in cases of inhibitory effects). If the postsynaptic cell reaches the threshold potential, it too will fire and so on throughout the neural network (Wilson 1999, Dayan & Abott 2001, Südhof 2013).

[6] The difference in electrical potential between the interior of a neuron and its surrounding exterior (Dayan & Abott 2001).
[7] about -70 mv.
[8] Exceptions exist. For instance, so-called 'leaky' channels tend to be open.

If this machinery of action potentials was subject to deterministic explanations, the problem of luck would hardly have the relevance it has today. However, this is not so. Consider the release of neurotransmitters into the synaptic cleft. When an action potential arrives at the presynaptic terminal, the chance of a single synaptic vesicle to release its contents can be as low as 20% (Auger & Marty 2000). In addition, the release of neurotransmitters from synaptic vesicles can happen even if no action potential arrives at the terminal at all (Dayan & Abott 2001). As a result of such observations, the release of neurotransmitters is commonly modelled as a stochastic process (Varela et al 1997, Tsodyks & Makram 1997, Abbott et al. 1997, Dayan & Abott 2001).

Another place in the process where we find probabilities is at the opening and closing of ion gates. Ion channels are commonly modelled as Markov Models, which are paradigmatic cases of stochastic models (Colquhoun and Hawkes 1995, Werndl 2016). On this view, channels move from open to closed states with a certain probability. I.e. at any given moment, there is a certain probability that the channel will change state (Dangerfield et al. 2010, Frank & Toshinori 2015). This stochastic aspect of ion gates has implications for further neuroscientific phenomena and have been shown to affect both the spike probability and the spike timing of cells and in turn, the computations carried out by neurons (Schneidman et al 1998, Diba et al. 2004, Jacobson et al. 2005, Dorval 2005, Carelli et al. 2005, Kole et al. 2006, Saarinen et al. 2008, Dudman & Nolan 2009).

The variability of synaptic transmission and the behavior of ion channels are just two sources of neural 'noise'. Noise is commonly understood as a kind of variability that results from random or unpredictable fluctuations (Faisal et al. 2008, McDonnell & Ward 2011). Nowadays, it is understood that noise, even noise at the microscopic scale, has important implications for neural activity at the higher 'macroscopic' levels (Harris & Wolpert 1998, Faisal 2008, Mcdonnell & Ward 2011, Uddin 2020). One important role of noise concerns its role in decision-making (Hanes & Schall 1996, Hesselmann et al. 2008). Voluntary self-initiated movements are preceded by a gradual build-up of neuronal activity (the readiness potential) (Kornhuber & Deecke 1965, Libet et al. 1983). This kind of gradual increase in neural activity preceding spontaneous movements has been found to be a common phenomenon (Romo & Shultz 1987, Kagaya & Takahata 2010, Fried et al 2011, Schurger et al. 2012,). For instance, the spontaneous decision of 'when' to move i.e. the decision to move *now*, corresponds to a threshold-crossing neural event that can be modelled in terms of a "leaky stochastic accumulator"[9] (Schurger et al. 2012, p. E2904). On this model, ongoing neural noise is continuously bringing the neural activity closer

[9] This has been shown in contexts with no specific temporal cue and only a general imperative to produce a movement at an unspecified time (Schurger et al. 2012).

to or farther away from the threshold for acting. When the noise accumulates enough to cross the threshold, a decision to move 'now' occurs and spontaneous voluntary action takes place. On this account, "the *precise* moment […] is determined stochastically by ongoing spontaneous fluctuations in neural activity" (Schurger et al. 2012, p. E2909). Also then at the level of neural activity underlying decisions and voluntary actions, do we find neural noise and innate variability play a role.

Here we see just three examples of the probabilities in neuroscientific explanations: In modelling synaptic transmission, in modelling the behavior of ion gates, and in modelling the neural activity underlying volitional decision-making. But exactly how such probabilities should be interpreted is a central question in contemporary philosophy. The three primary positions are the *frequency interpretation*, *Bayesianism*, and the *propensity theoretical* interpretation. Frequentism understands probability simply as the relative frequency of an event over time (Hájek 2019). It is principally uncommitted to any underlying metaphysics (Kendall 1949). Bayesianism is the 'subjective' interpretation of probabilities, which interpret probabilities epistemically as reflecting *degrees of belief* (de Finetti 1974, Hartmann & Sprenger 2010, Talbott 2016). The propensity theoretical interpretation interprets probabilities as physical *dispositions*, something objective and physically real similar to physical forces (Pierce 1931, Popper 1959, Hacking 1965, Mellor 1971, Gillies 1973, Kyberg 1974). On this view, chance is an ontological property of the world. As we will see in chapter 4, it will be argued that we have good reasons to accept a propensity theoretical interpretation of probabilities.

1.3 Folk-conceptions and experimental philosophy: Why should we care?

Intuitions play a core role in philosophy. From Gettier cases (Gettier 1963) to Frankfurt examples (Frankfurt 1969) to Mary's room (Jackson 1982, 1986) and the veil of ignorance (Rawls 1971) many of philosophy's most striking arguments and thought experiments derive their force from some – supposedly shared – intuitions. Indeed any argumentation must take *something* for granted for the conversation to take off at all. Any reason-giving game ultimately rests on a foundation that is itself 'groundless' (Wittgenstein 1969). When such foundational intuitions and presuppositions are not shared, people are likely to talk past each other. Potential disagreements run the risk of being no more than semantic differences about the definitions of words. When the axiomatic presuppositions differ between people, how can one starting point then be justified over another? In the free will literature, a common approach is to appeal to *folk-intuitions*. Philosophers often appeal either directly to the intuitiveness of one particular conception of free will over another (Wolf 1990, Kane 1999, Ekstrom 2002, Pink 2004, Baumeister 2008), or their arguments themselves appeal to the intuitive plausibility of

the principles from which the conclusions are derived (Frankfurt 1969, Van Inwagen 1983, Strawson 1986, Ravizza 1994, Fischer and Ravizza 1998). Commonly, appeals to folk-intuitions are made from 'the armchair' by philosophers who tend to assume that 'the folk' share whichever intuitions the given philosopher have themselves. Unsurprisingly, this has led to conflicting ideas about what exactly ordinary people's intuitions are.

This appeal to ordinary intuitions, and the conflicting views about them, has in turn resulted in the new philosophical research program of *experimental philosophy*. What constitutes *ordinary* intuitions is an empirical question that can be examined experimentally. Experimental philosophy conducts experimental studies of people's ordinary intuitions and judgments to investigate just such questions about ordinary intuitions (Knobe 2007, Nadelhoffer & Nahmias 2007, Alexander & Weinberg 2007). Experimental philosophy covers areas in epistemology, action theory, free will, and moral responsibility amongst others (Nadelhoffer & Nahmias 2007, Knobe 2007, Knobe & Nichols 2017). Part of the work is focused on 'experimental analysis' (Nadelhoffer & Nahmias 2007, see also 'proper foundation view' Alexander & Weinberg 2007), where the intuitions and common-sense concepts of ordinary people are investigated. The aim is to determine *what* the laypeople's intuitions and concepts really are. A second part of the work is focused on 'experimental descriptivism' (Nadelhoffer & Nahmias 2007), where the focus is on *how* these general intuitions are generated and what kind of cognitive processes drive particular intuitions of the general public.

The approach of experimental philosophy is not entirely uncontroversial, however. One objection is that studying folk-intuitions is irrelevant because the kind of studies that investigate these only ever get to *surface intuitions* and not *robust* intuitions (Kauppinen 2007). The idea is that the theoretical and well-trained intuitions of philosophers can be genuinely useful, but that the intuitions of ordinary people about philosophical matters are unreliable and irrelevant for any philosophical matters. As Nadelhoffer and Nahmias point out, this objection should offer an account for why philosophers' intuitions are more reliable that does not beg the question (Nadelhoffer & Nahmias 2007, Alexander & Weinberg 2007). It is easy to assume that one's own intuitions are reliable and that those who differ are not, but there is nothing to suggest that philosophers are any less prone to the biases of human psychology. Furthermore, in addition to the ordinary trap doors of human cognition philosophers are also at the risk of so-called 'theory contamination' (Goldman & Pust 1998).

A second, more common, objection is to ask why we should care about ordinary people's intuitions at all. The philosophically interesting question concerns the *truth* about free will, not the

ordinary person's judgment about it. Scientists and mathematicians don't care about the ordinary person's understanding of their subject matter – why should philosophers?

One reason is that, unlike scientists and mathematicians, philosophers investigate their topic by making use of intuitions (Sommers 2010, Pust 2019). Consider the following example of philosophical reasoning:

> "The degree to which Haji's conclusions are unintuitive must be weighed against how unintuitive it is to reject one or more of his premises. If a theory that uses certain principles as premises has components derived from these principles that are unintuitive, the principles would thereby be to some degree disconfirmed. It would be implausible to claim that in a theory that uses K and OW as premises, these principles would have a justificatory status so strong that it immunizes them against disconfirming pressures from their unintuitive consequences. This is true even if OW and K are conceived as axioms in a moral theory. One might begin with principles such as K and OW, and regard them as axioms because they are intuitively true and because they appear central to the theory. If the components of the theory derived from these principles conform to our intuitions, that would provide theoretical support for them. But if such derived components do not conform to our intuitions, that would to some extent disconfirm these principles" (Pereboom 2001, p. 144)

Here, Pereboom makes explicit a common trait of philosophical reasoning that is often applied implicitly: If the consequences of a set of principles in question conform to our intuitions this counts as theoretical support in favor of these principles. If the consequences of a set of principles are unintuitive, this counts – to some degree – as disconfirming the principles. In this way, intuitions play a different role in philosophy than they do in science and mathematics. In the latter disciplines, they are important for *generating* hypotheses and positions but do not play the same role in *justifying* positions (but if we ask mathematicians to justify their axioms they might also appeal to intuitions and ideas about what is self-evident). While traditional philosophers are happy to judge certain premises and conclusions as intuitive or unintuitive from the comfort of the armchair, on the basis of their own introspection and informal dialogues (usually with other philosophers or philosophy students), experimental philosophers take an empirical approach to evaluating the intuitiveness of the positions to be judged by the scale of intuitive conformity.

A second reason to look at ordinary intuitions is that it is our everyday notions that are tied to general interests about our self-conceptions and social practices: People want to know if they have free

will or knowledge in the ordinary sense of these words. Though common-sense *theories* are unlikely to be successful, philosophical theories should at least be partly constrained by common-sense *concepts* if they hope to satisfy the curiosities that spark interest to begin with. Theories completely detached from the ordinary meanings of these words risk "having nothing more than a philosophical fiction as its subject matter" (Mele 2001, p. 27). Philosophers should strive to not inadvertently 'change the subject' when using natural language expressions, especially when speaking about things that are of general interest such as the question of free will. If philosophers' work concerns only terminological artifacts, one is left wondering why anyone should care about the work of philosophers to begin with. As Kauppinen puts it: "Why should anybody care about what philosophers do if they just argued about their own inventions?" (Kauppinen 2007, p. 96). Perhaps some philosophers don't mind a cleft of apathy between their own work and the rest of the world, but for those of us who do, the ordinary concepts at the root of general interests serve as a relevant starting point for our inquiry. Certainly there is still room for revisionism and conceptual development - reflections should not only be constrained by intuitions but also be used to correct them. But revisionism needs extra justifications. For example by showing that folk-conceptions are confused or too imprecise for one's purpose (Kauppinen 2007). Therefore, let's start with the folk.

One of the foundational papers in the field is Nichols' and Knobe's paper from 2007 (Nichols & Knobe 2007). There, participants were presented with two vignettes. One described a deterministic universe where everything is completely caused by what went before it so that when an agent decides to have French fries, "it *had to happen* that [he] would decide to have French Fries" (Nichols & Knobe 2007, p. 669). The other described an indeterministic universe where human decision-making is not completely caused by what went before it so that it "*did **not** have to happen*" (Nichols & Knobe 2007, p. 669) that an agent at a given time decided exactly as he did. After having read the vignettes, participants were divided into two conditions: A concrete condition and an abstract condition. In the abstract condition, participants were simply asked if it was possible for a person in the deterministic universe to be fully morally responsible[10]. In the concrete condition, on the other hand, the question was designed to invoke a greater emotional response. The results showed that in the concrete condition a majority of participants (72%) gave the compatibilist answer (yes), while in the abstract condition a majority (86%) gave the *incompatibilist* answer (no). In other words, people seemed to apply different concepts of free will based on whether they were in the abstract or concrete condition. Nichols' and Knobe's interpretation was that people had genuine incompatibilist intuitions, but that the affective component

[10] In the early days, questions about moral responsibility was used as a proxy for intuitions about free will.

invoked by the grizzly details of the concrete conditions made participants unable to properly apply the incompatibilist concept in their judgment (the affective error hypothesis) (Nichols & Knobe 2007). They further corroborated this interpretation by demonstrating that people were more likely to give compatibilist answers in response to high-affect scenarios (64%) than in response to low-affect scenarios (23%). When these high- and low-affect scenarios were described in an indeterministic universe, people gave high free will ratings regardless of the affective salience.

In contrast to Nichols' and Knobe's paper, Nahmias et al. had previously conducted a similar vignette study, with different results (Nahmias et al. 2005). They employed two different vignettes. One was a scenario where a supercomputer could perfectly predict the actions of an agent by looking at the current state of the world (the super-computer scenario). The other was a scenario where time is turned back, and the universe is identically re-created over and over again with the agent acting the same way in every iteration (the roll-back scenario). What they found was that participants overall gave compatibilist answers (76%). Interestingly, Nahmias et al. checked to see if the compatibilist answers were driven by emotional reactions. They did so by employing scenarios where the agent performed blameworthy, praiseworthy, and neutral actions. Their findings revealed that in all cases the majority of participants gave compatibilist answers (76%, 68%, and 79% respectively). These results contradict the affective error hypothesis, as it cannot explain why people would give compatibilist answers in scenarios with low emotional valence, such as when the agent decides to go jogging.

These are puzzling findings. Why did Nahmias et al.'s vignettes generate opposite results to Nichol's and Knobe's vignettes? What is particularly puzzling is that the studies individually have clear results, but these results are inconsistent with each other. If they showed mixed results on their own, this would indicate that laypeople simply have mixed concepts. What they show instead is that laypeople overwhelmingly apply one particular concept, but which concept this is differs depending on which study we read – despite the studies applying similar methods. Since these early studies, some severe methodological challenges have been demonstrated, that helps shed light on these inconsistent findings. These are the methodological issues known as the *bypassing effect* and the *intrusion effect*. These effects are two different ways in which people misunderstand determinism: The bypassing effect drives apparent incompatibilist answers and the intrusion effect drives apparent compatibilist answers.

The bypassing effect is when people misinterpret deterministic descriptions to entail that the agent's actions are caused by forces that bypass the agent's conscious self. In other words, they wrongly interpret determinism to entail *epiphenomenalism*. For instance, in 2010 Nahmias and Murray reused the vignettes previously applied in the literature and asked participants if the agent's decisions, wants,

beliefs, and past had an effect on what the agent ended up doing and if the agent had control over what they do (Nahmias & Murray 2010, Murray & Nahmias 2012). What they found was that people frequently said that an agent's decisions, wants, and beliefs *did not* have an effect on the agent's behaviour in deterministic universes. This clearly demonstrates an *epiphenomenal* reading of determinism. They also found an inverse relationship between free will and moral responsibility scores and bypassing scores. In other words, there was a strong correlation between responses that said an agent lacked free will and epiphenomenal readings of deterministic vignettes. Notably, they also showed that the bypassing effect was more prominent in abstract than in concrete conditions, and more common in Nichols' and Knobe's vignettes than in Nahmias et al.'s vignettes. Their results demonstrated that a majority of incompatibilist responses in previous studies were driven by this bypassing effect (~80%) (Nahmias & Murray 2010). When people responded that free will was impossible in deterministic universes, it was because they mistakenly thought mental causation was impossible in such a universe. These findings could then help explain why Nahmias et al.'s study and Nichols' & Knobe's study had garnered conflicting results: Because the vignettes in Nichols' and Knobe's study were more prone to invoke the bypassing effect, which led to incompatibilist results.

When participants give incompatibilist answers, it is then not an indication that they apply the incompatibilist concept of free will, since they are not responses "about the incompatibility of determinism, understood in the way relevant to the philosophical debates, and free will" (Murray & Nahmias 2012, p. 15). To read determinism as entailing epiphenomenalism is an interpretation in stark contrast and at odds with the philosophical concept of determinism. As previously noted, most accounts of free will agree that it requires *mental causation*. When participants in these earlier studies give *apparent* incompatibilist answers, these answers are generated by participants understanding the deterministic descriptions differently than as philosophers intend. We might think, then, that when people do *not* misunderstand deterministic vignettes, they give compatibilist responses – or at least this would seem so, were it not for the *intrusion effect*.

The intrusion effect is when people fail to track the deterministic component of the description and import indeterministic metaphysics into the scenarios: They do not properly grasp the nature of necessity that characterizes determinism (Rose et al. 2015). For instance, Nadelhoffer et al. tested the presence of intrusion by presenting participants with deterministic vignettes and asking participants if it was possible for the agent in the scenario to change his mind, if there was a slight chance that he would act differently, or if it was possible for the agent to act differently at the time (Nadelhoffer et al. 2020). Answering yes to these questions indicates that the participant interprets the vignette to include

indeterministic metaphysics about the agent's decisions and actions. Since determinism entails that with the exact same initial conditions an agent *cannot* act differently than he or she does, answering yes to the Nadelhoffer et al. questions shows that participants do not properly track the implications of the deterministic descriptions. Their results also showed that intrusion correlates with free will ratings. In other words, when people misinterpret determinism to include indeterministic metaphysics they are more likely to respond that the agent had free will. This finding problematizes the results that seem to show that laypeople have a compatibilist concept of free will. If participants read deterministic vignettes to include indeterministic metaphysics, their replies do not really reflect compatibility between determinism and free will – but between indeterminism and free will. Just as the bypassing effect problematizes findings that show apparent incompatibilist folk-conceptions, the intrusion effect problematizes findings that show apparent compatibilist folk-conceptions.

Since the days of Nichol's and Knobe's and Nahmias et al.'s foundational papers, no consensus about the folk-concept of free will has emerged. In large part because of the above noted methodological challenges. As our starting point will be to try and establish the common-sense meaning of 'free will', we will in the next chapter try to settle this discrepancy by looking at the folk-conception of free will. This will be done by applying similar vignettes as those used previously in the literature but while controlling for both the intrusion effect and the bypassing effect. The outcome of investigating folk-conceptions of free will, will serve as the starting point and default concept of free will that the rest of the thesis builds upon.

1.4 Layout of the thesis

Collectively, the four papers that make up the chapters of this thesis aim to answer the primary research question - how probabilistic uncertainties in the mechanisms underlying intentional action should affect our understanding of free will - in the following way:

First, by asking what does 'free will' mean? Despite centuries of intense focus on the problem of free will, the expression 'free will' has even today no settled-upon technical definition. Sometimes it is defined with reference to moral responsibility, sometimes it is defined as the ability to do otherwise, and there is frequent talk of 'compatibilist' free will and 'libertarian' free will as two different things. When we have terminological anarchy, it is tempting to adopt an 'anything goes' attitude where people are free to define the terms as they please, so long as they define them explicitly. Here, I will instead take a less arbitrary approach. The expression 'free will' is already present in natural language, but what kind of free will is in play there? The first paper (chapter 2) applies the approach of experimental

philosophy and investigates which philosophical concept best approximates the folk-conception of free will. In particular, I investigate if the idea of probabilistic decision-making clashes- or aligns with the folk-conception of free will. What we find is that the folk-concept of free will is compatible with a probabilistic world but incompatible with a deterministic world. In other words, laypeople have an incompatibilist but not hard incompatibilist conception of free will. The second paper (chapter 3) expands upon the findings of the first experiment, by investigating if belief in physicalism or substance dualism affects laypeople's *concept* of free will. Do physicalist beliefs drive more incompatibilist and hard incompatibilist conceptions of free will? What we find is that belief in physicalism *appears* to drive incompatibilist free will conceptions, but that this appearance is driven by different understandings of descriptions of determinism. Specifically, physicalists tend to interpret determinism to entail epiphenomenalism. Our findings show that, when this understanding error is corrected for, we do not find any differences in free will concepts driven by differences in physicalist or substance-dualist beliefs.

The second part of the thesis asks: Is ontological indeterminism true? If free will – in the ordinary sense – requires ontological indeterminism we must find epistemic justifications for an indeterministic framework if talk of human agency as free is to be more than hopeful thinking. The third paper (chapter 4) asks how we should interpret the probabilities of our scientific models. Should we think of them as merely *epistemic* or do we have good reason to accept them as genuinely *ontological*? I consider three different epistemological approaches and argue that whichever approach we choose we have sufficient reason to accept ontological indeterminism in light of best-explanation probabilistic models from science. I thus argue that we can reasonably accept ontological indeterminism as true, also at the level of the brain and in the processes that underlie mental activity.

The third and final part asks how we must reconceptualize free will in light of the problem of luck. If indeterministic randomness in human decision-making poses a problem for free will – as the problem of luck illustrates - what kind of control might an indeterministic agent have to alleviate this problem? The fourth paper (chapter 5) builds upon the common-sense notion of free will. It is argued that the problem of luck is ultimately unsolvable. It confronts us with a conceptual choice: We must either abandon the idea of free will in indeterministic agents or we must accept some element of randomness as part of free will. Going with the second option, I propose that ontologically indeterministic agents must have a way to influence their own internal probabilities for their will to be free. If they are able to make self-directed decisions that target and alter the agent's own motivational structure, i.e. if they have the capacity for *recursive decision-making*, indeterministic agents gain an important kind of control over their internal randomness. With recursive decisions, an agent can then –

to a certain degree – influence the indeterministic randomness innate to their decision-making. It is further argued that because this influence is always a matter of degrees only free will and human agency is best understood as a graded phenomenon.

2. Free Will Properly Understood: Folk-Conceptions are Incompatibilist, But Not Hard Incompatibilist

Abstract

Could we be free in a world where everything is strictly determined? What if the world was governed by chance? Philosophers disagree about these two sides of the compatibility problem, including which answer is more intuitive. To capture this latter part, philosophers have turned to experiments, to investigate whether folk-conceptions of free will are compatible with determinism. One key problem of such studies is that it is unclear if participants' understanding of indeterminism reads as simply the absence of determinism or the concrete presence of probabilistic chances. Another common problem with experiments asking non-philosophers about free will and determinism is that laypeople misunderstand determinism: Either by inferring that it rules out mental causation (the bypassing effect) or by still injecting indeterminism into the description (the intrusion effect). In our experiment, using a within-subject design (N = 158), we control for intrusion, bypassing, and explore the understanding of probabilistic chance in relation to free will. We find that many participants struggled to understand the notion of probabilistic chance in free will. Additionally, we show that when people properly understand philosophically relevant concepts such as determinism, they deny that free will can exist in a deterministic universe but accept that free will is compatible with chance. As such, we can show with better clarity that laypeople are incompatibilists but not hard incompatibilists.

2.1 Introduction

The question about whether free will is an illusion or not is often rooted in the idea that there is some universal phenomenological experience or psychological intuition, which is shared by all or nearly all people across time and cultures. The assumption that free will is true serves a key role in justifying our social and legal practices, and in shaping our self-conceptions. When asking about the reality of free will, what we want to know is if our intuition about free will is accurate or not: Does it correspond to reality or is it a cognitive trick the mind plays upon itself? (Wegner 2002). A major obstacle in the way of gaining any consensus on the matter is that philosophers disagree about what exactly constitutes the

concept that is at the root of our questioning. What is the intuitive idea of free will which we should look for in reality? One thing most philosophers do agree upon is that free will is a particular kind of *mental causation*, where mental states and events play a key role in bringing about actions. But when it comes to establishing the particulars, the disagreements begin. Specifically, philosophers tend to disagree about whether free will can exist in a strictly deterministic universe or if it requires indeterminism. Determinism is the idea that causes and effects follow laws of strict necessity, so that "at any instant exactly one future is compatible with the state of the universe at that instant and the laws of nature" (Mele 2006, p. 3). Indeterminism is commonly defined as the absence of determinism. According to *incompatibilism* free will is incompatible with determinism: We cannot have free will if everything, including human decision-making and action, follows with strict necessity from what came before. According to c*ompatibilism*, on the other hand, free will is compatible with determinism. For compatibilists, it is unimportant whether human decisions are a consequence of a deterministic machinery or not. The important thing is that mental causation is real and that the agent is free from external force and manipulation. So long as the will is causally efficacious and the agent is not being manipulated, the agent has free will.

These two positions - compatibilism and incompatibilism - constitute the two primary schools of thought regarding free will. But recently a third notion has found proponents. *Hard incompatibilism* is the position that free will is incompatible both with the necessity of determinism and the chanciness of indeterminism (Pereboom 2001, 2014a, Pereboom and Caruso 2002, Caruso 2012, 2019b). The hard incompatibilist notion of free will is well captured by Kastrup when he writes that "we often think – incoherently – of free-willed choices as neither determined nor random" (Kastrup 2020). Occasionally, a fourth conception of free will finds mention (O'Connor & Christopher 2020): The concept of free will that *requires* determinism to be true, i.e. only if determinism is true is free will possible (the opposite of incompatibilism). Let's call this position *exclusive determinism*. Together, these four concepts logically exhaust the possible ways free will may depend upon or be excluded by determinism and indeterminism (fig. 1 B)[11].

The four competing free will concepts are then: 1) The incompatibilist notion of free will, which requires indeterministic metaphysics. 2) The compatibilist notion of free will, which has no strict metaphysical requirements about determinism or indeterminism. 3) The hard incompatibilist notion of free will, which requires that human decision-making is free from both metaphysical determinism and

[11] Agent-causal concepts of free will are omitted from this list as these are (commonly but not necessarily) a branch of incompatibilism.

indeterminism. 4) The exclusive determinist notion of free will, which requires deterministic metaphysics.

The logico-philosophical discussion about free will is at an impasse, stuck in "dialectical stalemates" (Murray & Nahmias 2012, p. 1) about which of these concepts deserves the name. This semantic tug of war has resulted in an unorthodox turn in philosophy of free will – it has made the intuitions of ordinary people a key focus.

2.2 The turn to folk-intuitions

To fix the object of investigation and root the endeavor in something psychologically relevant, many philosophers take as their starting point the concept of free will as it is commonly understood. As Mele puts it, a philosophical concept of free will detached from the common-sense understanding of the phenomenon "…runs the risk of having nothing more than a philosophical fiction as its subject matter." (Mele 2001, p. 27). If philosophy was "concerned only with the technical sense of the concept, it would be divorced from the concerns that led us to philosophical investigation of the concept in the first place…" (Alexander & Weinberg 2007, p. 58).

However, philosophers disagree about which concept of free will best suits the supposedly common-place intuition that first gave rise to the expression. Some philosophers claim that incompatibilism is most intuitive. That "Most ordinary people start out as natural incompatibilists" (Kane 1999, p. 218), that "we come to the table, nearly all of us, as pre-theoretical incompatibilists" (Ekstrom 2002, p. 310), and that "compatibilism is not something naturally believed, but something that has to be taught - by professional philosophers, in philosophy books, and through philosophy courses" (Pink 2004, p. 43). Others claim, on the other hand, that compatibilism is the most intuitive notion: That it "accord with and account for the whole set of our intuitions about responsibility" (Wolf 1990, p. 89) and that it corresponds "…to what laypersons generally mean when they distinguish free from unfree action" (Baumeister 2008, p. 14). These appeals to common-place intuitions are rooted in an attitude that the views that are consistent with folk-intuitions "have prima facie plausibility that the views that are inconsistent with folk intuitions do not" (Shultz et al. 2011, p. 1723).

These appeals to folk-intuitions have lead philosophers to empirically investigate what the ordinary concept of free will is. A common approach is to apply the contrastive vignette method, where participants are presented with two or more descriptions of hypothetical universes which operate by different rules and then asked to rate whether or not agents in such universes could possess free will

and be held morally responsible. However, four primary problems confront this research program. For one, it is not clear whether what is being tested is *intuitions* at all. It is rare to find reaction time checks or other methods such as the Cognitive Reflection Test (Frederick 2005) to control for how reflective or intuitive the answers participants provide are. What is called folk-intuitions in the literature may then more appropriately be called folk-*conceptions* of free will[12]. Secondly, most participants are undergraduate students (Nichols & Knobe 2007, de Brigard et al. 2008, Preston et al. 2013, Nahmias et al. 2014). If we want to understand the conceptions of the general public, research should be conducted on a representative sample that includes participants outside academia and across ages. Thirdly, the experiments conducted show varying and mixed results. Some studies seem to show that folk-conceptions fit the incompatibilist notion (Nichols 2004, 2006a, 2006b, Nichols & Knobe 2007, Sarkissian 2010), while others appear to show that the folk-concept fits the compatibilist idea of free will (Monterosso et al. 2005, Nahmias et al. 2005, Nahmias 2006, Nahmias et al. 2007, Nahmias & Murray 2010, Monroe & Malle 2010, Murray & Nahmias 2012, Monroe et al. 2014, Feltz & Cokely 2019). Because of this ambiguity, the semantic dispute remains unsettled: There is still no consensus on what the pre-theoretical folk-concept of free will is. The fourth and perhaps the biggest problem is that participants often misinterpret the key philosophical concepts philosophers are trying to convey. Specifically, they often misinterpret descriptions of determinism in two important ways known as *bypassing* and *intrusion*.

The bypassing effect is when people misinterpret descriptions of determinism to entail that the agent's actions are caused by forces that bypass the agent's conscious self. In other words, they wrongly interpret determinism to entail *epiphenomenalism*[13] (Nahmias & Murray 2010). There is a strong correlation between people misreading deterministic descriptions as entailing epiphenomenalism and responding that free will is impossible in such a universe. In other words, the bypassing effect has been found to drive apparent incompatibilist responses. 'Apparent' because the understanding of determinism that genuine incompatibilists think are at odds with free will does not exclude mental causation. The intrusion effect, on the other hand, is when participants fail to track the deterministic component of the descriptions of determinism and import indeterministic metaphysics into their interpretation (Nadelhoffer et al. 2020, Rose et al. 2015). When people misunderstand determinism in this way, they tend to ascribe free will to agents in such a universe. The intrusion effect thus drives

[12] For this reason, we will here adopt the language use of folk-conceptions instead of speaking of intuitions.
[13] Epiphenomenalism is the idea that mental states play no causal role in the world (Robinson 2019).

apparent compatibilist responses. 'Apparent' because the understanding of determinism that compatibilists think is compatible with free will does not include indeterminism at any level.

To get a proper picture of the folk-concept of free will, it is then important that studies are conducted on representative samples and include checks for both the bypassing and the intrusion effect to ensure that participants understand the key concepts in the philosophically relevant way.

2.3 Aim and Hypothesis

The general aim of this study is to contribute to the literature on folk-conceptions of free will in two ways, one theoretical and one methodological.

The theoretical aim is to account for all four of the possible free will profiles people may have concerning free will and determinism/indeterminism, instead of limiting the focus by assuming that laypeople are either compatibilist or incompatibilist. Previous studies have framed indeterministic universes in a negative framing (i.e. as *not* deterministic). This is problematic for two reasons. First of all, understanding indeterminism in such terms depends on properly understanding determinism in the first place and upon being able to derive a concrete concept from the negation of determinism. Decades of psychological research warn against possible side effects of using negative or loss framings such as 'there is no x' when testing people's judgments and behavior (Lyengar 1990). It then seems sound to check whether responses would be different if the 'non-deterministic' universe is described as the absence of deterministic laws or as the presence of chance events (probabilistic laws). To avoid this framing problem we expand upon the existing literature by including a probabilistic description of indeterminism[14]. Secondly, when indeterminism is framed as the negation of determinism, replying that free will is possible in indeterministic worlds just reflects a concept of free will that is incompatible with determinism: An aspect that both incompatibilism and hard incompatibilism share. By examining if the folk-concept of free will is compatible with a probabilistic universe, we accommodate the possibility that laypeople are hard incompatibilists. In the free will debate, things of a probabilistic nature are considered random (Van Inwagen 1983, 2017, Strawson 1986, Kane 2002, Mele 2006, Roskies 2006, 2010, Haggard 2010). Following Kastrup, we expected that folk-conceptions of free will are at odds with both the ideas of randomness and determinism in human decision-making. Our theoretical hypothesis was then that the folk-concept of free will is hard incompatibilist: That 1) people would not ascribe free

[14] In the context of free will and ontology, the terms 'indeterminism', 'chance', 'randomness', and 'probabilistic' are commonly used interchangeably. This use is adopted for the purpose of this paper.

will to agents in deterministic worlds and 2) that when indeterministic worlds are described in a probabilistic framing people would generally not ascribe free will to agents in such worlds.

Additionally, the methodological aim is to accommodate the previously mentioned problems that have been demonstrated in the literature – especially to limit and control for both the bypassing and intrusion effects. Nichols' and Knobe's vignettes have become standard use in the literature (Nichols & Knobe 2007, Feltz et al. 2009, Nahmias & Murray 2010, Cova et al. 2012, Nichols & Rose 2013, Feltz & Cova 2014), but in their descriptions, they put special weight on human decision making as "the one exception" (Nichols & Knobe 2007, p. 669) in an otherwise deterministic universe. We suspect that this formulation causes the bypassing effect, by making participants interpret the vignette in terms of mental causation contra non-mental causation (epiphenomenalism). To limit the presence of bypassing we included explicit descriptions of mental causation in both deterministic and indeterministic vignettes (see section 2.4). In addition, we include controls for both the bypassing and intrusion effects. Our methodological hypothesis was that including explicit descriptions of mental causation in the vignettes would reduce the presence of bypassing so a majority of participants would avoid the bypassing effect.

2.4. Methods and materials

An online survey of 158 respondents was conducted (79 male, 78 female, 1 other[15]). Participants' age ranged from 18-61 years with a mean age of 30.8. An a priori t-test using G*Power (Alpha = 0.05, power = 0.95, effect size = 0.8) was used to determine an estimated sample size of 84. Because one of the main aims of the study was to check for responses that *weren't* cases of either bypassing or intrusion, and due to the high presence of bypassing and intrusion effects, a total of 173 participants were recruited. Of these 173 participants 15 were excluded: 3 were excluded for being duplicate participants, 1 was excluded for responding in a non-English language and 6 were excluded because a software error caused them to be presented with the same vignette twice in a row. Because the aim was to investigate the conceptions of *laypeople* i.e. people who do not have a background in academic philosophy, 4 participants were excluded for having a Bachelor's degree in philosophy and 1 was excluded for having a Master's degree or higher in philosophy. All participants excluded by these measures were paid for their participation. The final sample size after exclusions was then 158 participants.

The experiment was built and run using *Qualtrics*. Participants were recruited using Amazon Mechanical Turk (Mturk, *n* = 73) and Prolific Academic (PA, *n* = 90). As Mturk has mostly American users,

[15] These were the three options available for participants.

and PA has mostly UK users, using both services gives a more diverse population of participants. The experiment lasted 11 minutes on average. Participants were paid 1.5€ (Mturk) or £1 (PA) respectively, corresponding to an hourly wage of approximately 7.5€. To exclude bots, participants on Mturk were filtered to only include users with more than 1000 completed assignments (HITs) and with a HIT approval rating of >99%. Because piloting of the PA users did not reveal any bots, no extra measures were taken to filter participants from this platform at this point in the recruitment process. Ethical approval was granted by the ethics committee of the School of Advanced Study at the University of London (SASREC_1819-313C-R).

For the purpose of this experiment, six new vignettes were designed. Three different descriptions of universes described by so-called 'rollback scenarios' (Nahmias et al. 2005, Nadelhoffer et al. 2020), each with two types of decisions: A high stake and a low stake decision. The three kinds of universes were deterministic, indeterministic, and probabilistic. Although both the indeterministic and the probabilistic descriptions describe ontologically indeterministic worlds, the descriptions differ in the following way: The description that we here call 'indeterministic' mirrors the descriptions that have been used previously in the literature, whereas the 'probabilistic' description describes an indeterministic world such that the probabilistic aspect of indeterminism is made salient to the reader. The two types of decisions the agents in the vignettes make are to decide to marry their fiancé (high stakes) or to have fries for dinner (low stakes). The difference in stakes was included to see if people applied different concepts based on whether the decision was habitual and inconsequential (choosing fries) compared to grander deliberate life-altering decisions (marrying one's girlfriend). Below are the low-stake versions of each vignette.

<u>Deterministic vignette (low-stake)</u>

"Imagine a universe (**Universe A)** in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present.

In this universe human decision making is caused by mental states such as beliefs, desires and intentions. These beliefs, desires and intentions are still completely caused by what came before them. So that if everything in the past from the beginning of the universe was exactly the same, then every belief, desire and intention that a person has in this universe had to occur exactly as they did.

For example one day Louis decides to have French Fries for lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until Louis made his decision, then it had to happen that Louis would decide to have French Fries at that point in time. If time was turned back over and over again, to the exact same moment just before Louis made his decision, then it had to happen that he would decide to have French Fries at that moment every time."

Indeterministic vignette (low-stake)

"Imagine a universe (**Universe B)** in which everything that happens is *not completely* caused by whatever happened before it. This is true from the very beginning of the universe, so what happens in the present of the universe is not *completely* caused by what happened before it, and so on into the future.

In this universe human decision making is influenced by mental states such as beliefs, desires and intentions. These beliefs, desires and intentions are still not completely caused by what came before them. So that even if everything in the past from the beginning of the universe was exactly the same, then every belief, desire and intention that a person has in this universe did not *have to* occur as they did.

For example one day Karl decides to have French Fries for lunch. Like everything else, this decision was not completely caused by what happened before it. So, if everything in this universe was exactly the same up until Karl made his decision, then it did not *have to happen* that Karl would decide to have French Fries. If time was turned back over and over to the exact same moment just before Karl made his decision, then it did not have to happen that he would decide to have French Fries at that moment every time."

Probabilistic vignette (low-stake)

"Imagine a universe (**Universe C)** in which everything that happens *is a matter of probabilities*, with no potential outcome having a 100% of a 0% chance of happening. This is true from the very beginning of the universe, so that what happens in the present of the universe is a matter of probabilities, and so on into the future.

In this universe human decision making is influenced by mental states such as beliefs, desires and intentions. These beliefs, desires and intentions are still a matter of probabilities. So that even if everything in the past from the beginning of the universe was exactly the same, then every belief, desire and intention that a person has in this universe *had a specific probability of occurring* as they did.

For example one day Peter decides to have French Fries for lunch. Like everything else, this decision was a matter of probabilities. So, if everything in this universe was exactly the same up until Peter made his decision, then *there was a certain probability* that Peter would decide to have French Fries at that point in time. If time was turned back over and over again, to the exact same moment just before Peter made his decision, then some of the times he would decide to have French Fries at that moment, and some of the times he would decide to have salad at that moment."

For the high-stake descriptions, the last paragraph was exchanged with a similarly worded paragraph describing the agent deciding to marry his girlfriend (see supplementary notes 1). All of the scenarios were deliberately designed to be non-moral scenarios (see section 2.6 for a discussion of this).

The experiment consisted of a within-subject design. Each participant read a consent form, a set of instructions, and was assigned at random to one of two experimental conditions (Fig. 1, A). In both conditions, participants were presented with two vignettes, presented in a random order. In condition 1 (determinism-indeterminism) participants were presented with one deterministic vignette and one indeterministic vignette. In condition 2 (determinism-probability) participants were presented with one deterministic vignette and one probabilistic vignette. Each vignette randomly included either a high stake or a low stake decision.

After reading the first vignette, participants were asked: "In Universe [x], is it possible to have free will?". Responses were recorded on a 7-point Likert-type scale (from -3: 'No' to 3: 'Yes', with the intermediate 0: 'Don't Know'. For a discussion of this, see section 2.6). Participants were then asked to provide reasons for their answers. After this, they had to answer a list of understanding questions before moving on to the second vignette, where the procedure was repeated. The understanding questions were questions about whether or not mental states had causal effects in the described universe and whether or not the agent could have done otherwise (see supplementary notes 2). By presenting participants with 2 vignettes, we were able to map out their fully-fledged free will profiles.

For statistical analysis of graded data, student's t-test was used when sample sizes are equal and Welch's t-test was used when sample sizes are unequal. For the binary coded data, logistic regression was used to analyze the data.

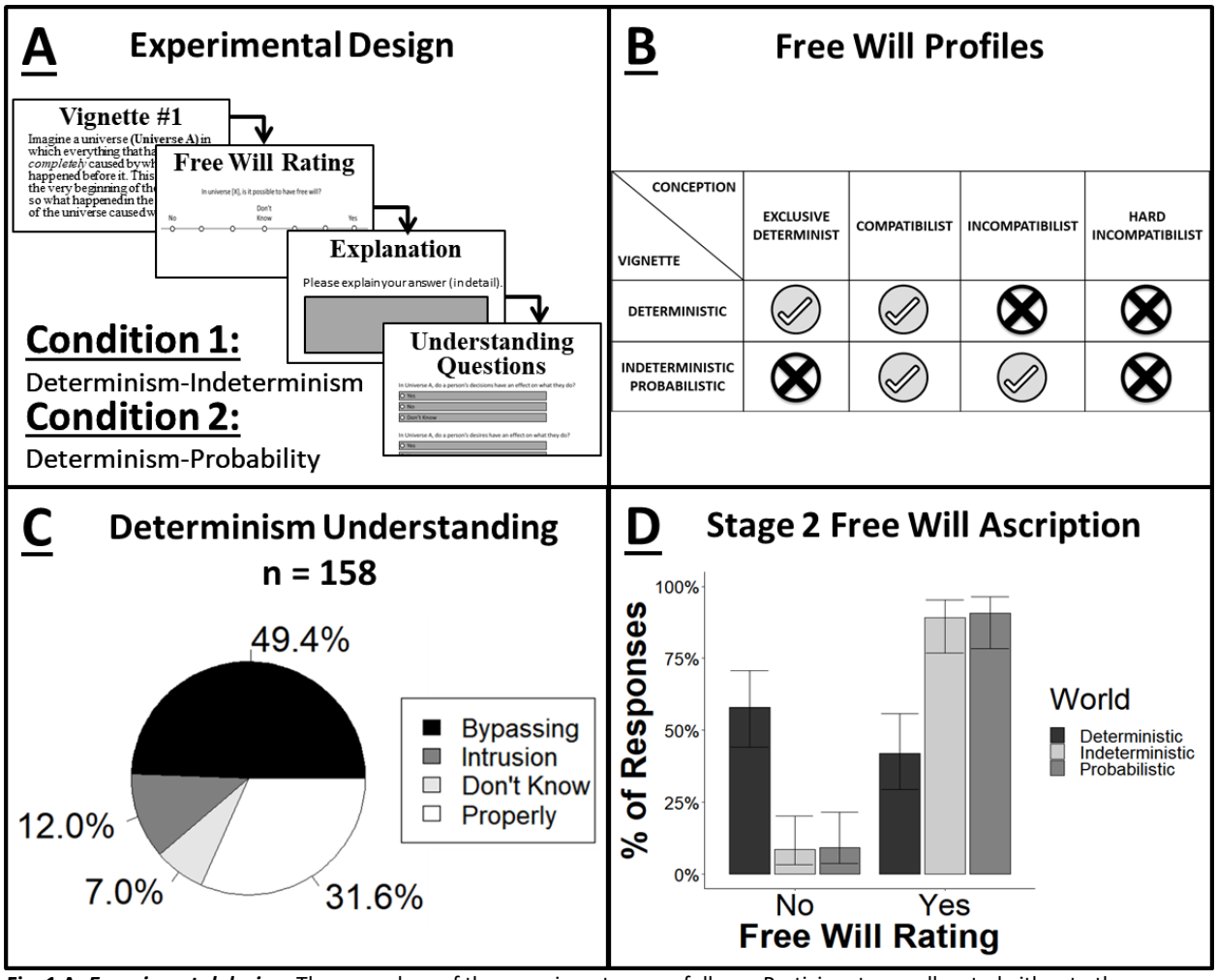


Fig. 1 A. ***Experimental design*****.** The procedure of the experiment was as follows: Participants are allocated either to the determinism-indeterminism condition or the determinism-probability condition. After that, they are presented with 1 vignette chosen at random and then asked to rate if it is possible to have free will in the described universe. They are then asked to explain their rating and asked understanding questions. Afterwards, they are presented with the 2nd vignette and the procedure is repeated. **B.** ***Response pattern of free will profiles*****.** Show how free will responses to the different vignettes map onto the different concepts of free will. **C.** ***Determinism Understanding***. Show the percentage of participants that committed bypassing, intrusion, responded 'don't know', and properly understood the deterministic vignette. **D.** ***Stage 2 Free Will Ascriptions***. Shows the free will ascription to the three different descriptions (deterministic, indeterministic, and probabilistic) when no understanding errors are present. (This graph omits 1 response in the 'don't know' category). Error bars reflect confidence intervals.

2.5 Results

66 participants were assigned to the determinism-indeterminism condition, 92 participants were assigned to the determinism-probability condition. The different number of people in the two conditions

was due to the varying presence of understanding errors[16]. While collecting data, partial analysis was run to check for the presence of bypassing and intrusion effects without analyzing free will ratings. Because understanding errors were more present in the determinism- probability condition, it was necessary to recruit extra participants for this condition to reach the minimum goal of at least a 42/42 split of responses with no understanding errors present. This reflects the fact that participants had a harder time understanding the vignettes in the determinism-probabilistic condition than in the determinism-indeterminism condition (for a discussion of this, see section 2.6).

To evaluate the effects of bypassing and intrusion, we processed the data in three different screening stages: Stage **(1)** included all responses to vignettes without accounting for understanding errors, as is commonly done in previous studies (Monterosso et al. 2005, Nahmias et al. 2005, Nichols & Knobe 2007, Sarkissian 2010). Stage **(2)** included only responses to vignettes where the metaphysics had been sufficiently well understood. Understanding questions were scored and responses where an understanding error was detected were excluded from the analysis (see supplementary notes 3). Stage **(3)** included only the responses of those participants who had properly understood the metaphysics of **both** vignettes they were presented with.

At every stage the data was processed in two ways: As graded data as collected via Likert-like scale responses, and as binary data (-3 to -1 coded as 'no', 1 to 3 coded as 'yes' and '0' coded as don't know). An order-effect was found with the probabilistic vignette at stage 1 of the analysis (paired $t(67) = -2.1$, $p < .05$) but it disappeared in stages 2 and 3. At no point in the analysis did we find any effect of stake or any effect of response-time (measured with reaction times).

Stage 1: All responses without understanding restrictions

When not controlling for understanding errors, 31.7% of participants ascribed free will to agents in deterministic worlds ($N = 158$, $M = -1.10$, SD = 2.09). 89.3% ascribed free will to agents in indeterministic worlds ($n = 66$, $M = 2.2$, SD = 1.50). And 61.9% ascribed free will to agents in probabilistic worlds ($n = 92$, $M = 0.81$, SD = 2.13) (Fig. 2, A and B).

Replies from participants in the determinism-indeterminism condition were significantly more prone to ascribing free will to indeterministic worlds than to deterministic worlds (paired $t(65) = -8.5$, $p < .001$; logistic regression, $p < .001$). Similarly, replies from participants in the determinism- probability condition also showed a significant difference in free will attribution, with a lower score in deterministic worlds (paired $t(91) = -6.6$, $p < .001$; logistic regression, $p < .001$). Importantly, when comparing free will

[16] Bypassing, intrusion, and conditionalizing (see section 2.6)

ratings to the indeterminism vignettes from the determinism-indeterminism condition with free will ratings to the probabilistic vignettes in the determinism- probability condition, a significant difference was also found, with free will ascribed less often to agents in probabilistic worlds (Welch t(156) = 4.8, $p < .001$; logistic regression, $p < .001$).

Stage 2: Responses when excluding misunderstandings

Of the participants with at least 1 response that passed the understanding check, 24 were male, 25 female, and 1 other. Their age ranged from 18-54 years with a mean age of 29.3 years. When controlling for understanding errors we were left with 50 responses to deterministic vignettes (~68% loss) (Fig. 1, C), 46 responses to the indeterministic vignettes (~30% loss), and 43 responses to probabilistic vignettes (~53% loss).

Among these replies, 42.0% rated the agent to have free will in deterministic worlds ($N = 50$, $M = -0.5$, SD = 2.30), 89.1% rated agents to have free will in indeterministic worlds ($N = 46$, $M = 2.2$, SD = 1.40) and 90.7% rated agents to have free will in probabilistic worlds ($N = 43$, $M = 2.0$, SD = 1.30) (Fig. 2, C and D).

Comparing the responses from the determinism-indeterminism condition again showed significantly higher free will scores in response to indeterministic worlds than to deterministic worlds (Welch t(82) = -7.0, $p < .001$; logistic regression, $p < .001$). Similarly, responses from the determinism-probability condition also showed a significant difference in free will attribution, with a lower score in deterministic worlds (Welch t(79) = -6.6, $p < .001$; logistic regression, $p < .001$). However, when comparing free will ratings from the indeterministic vignettes to the free will ratings from the probabilistic vignettes, the difference between these free will ratings was no longer significant (Welch t(87) = 0.6, $p > .05$; logistic regression, $p > .05$).

Stage 3: Complete free will profiles of participants

To get a picture of the fully-fledged free will profiles of participants, an analysis was run on responses from participants who understood *both* vignettes they were presented with in the philosophically relevant ways ($n = 31$, 15 male, 15 female, 1 other). Their age ranged from 20-48 years, with a mean age of 28.5 years.

Among these, 28.9% rated the agent to have free will in deterministic worlds ($n = 31$, $M = -1.16$, SD = 2.06), 81.25% rated agents to have free will in indeterministic worlds ($n = 16$, $M = 1.9$, SD = 2.01) and 93.3% rated agents to have free will in probabilistic worlds ($n = 15$, $M = 2.53$, SD = 1.06). Having

participants who properly understood both vignettes gives us the opportunity to look at their full free will profiles when no understanding errors take place. Of these 31 participants, 2 had hard incompatibilist conceptions (6.5%), 20 had incompatibilist conceptions (64.5%), 7 had compatibilist conceptions (22.5%) and 2 had exclusive determinists conceptions (6.5%).

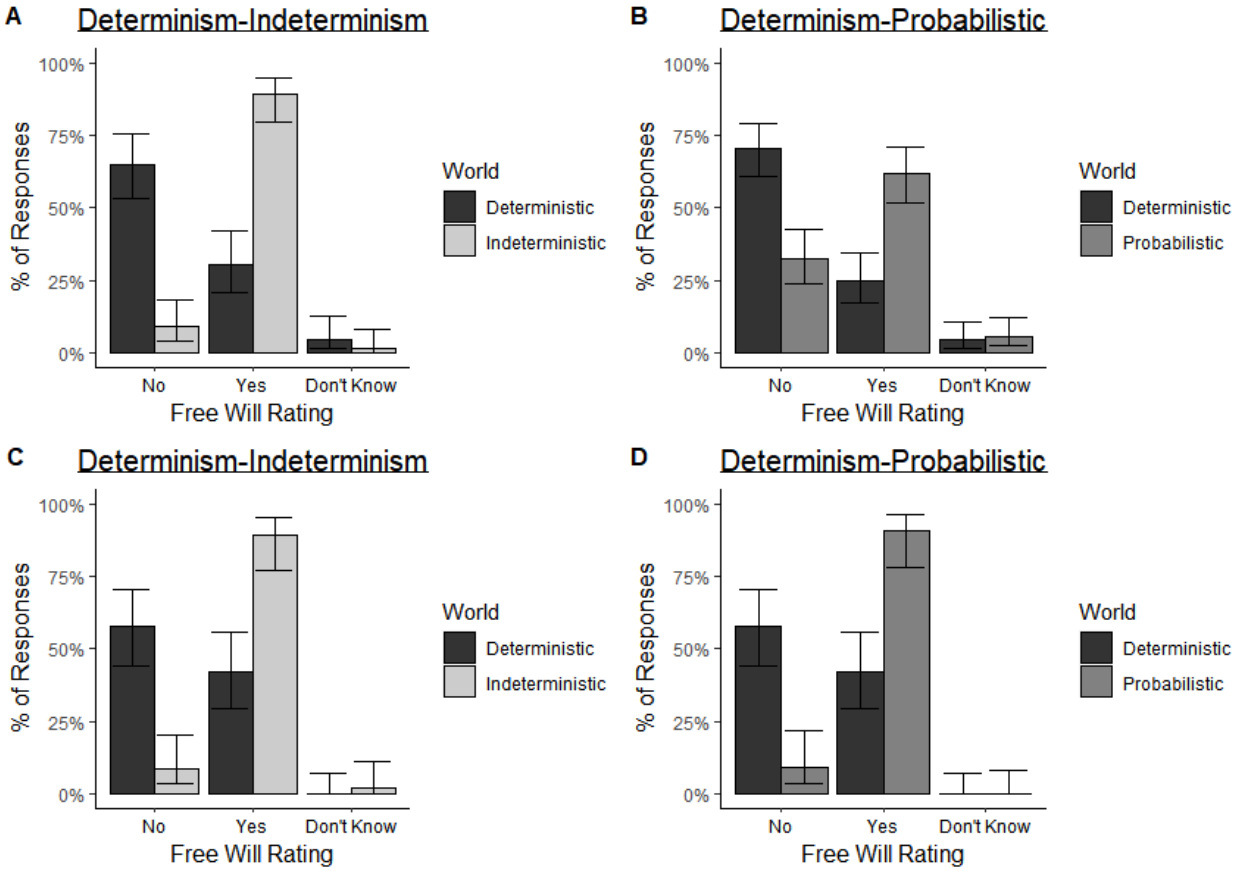


Fig. 2 Results as binary data. **A.** Shows the distribution of responses from the determinism-indeterminism condition at stage 1 of the analysis. **B.** Shows the distribution of responses from the determinism-probability condition at stage 1 of the analysis. **C.** Shows the distribution of responses from the determinism-indeterminism condition at stage 2 of the analysis. **D.** Shows the distribution of responses from the determinism-probability condition at stage 2 of the analysis. Error bars reflect confidence intervals.

2.6 Discussion

Our study showed two primary results. One related to the theoretical hypothesis and one related to the methodological hypothesis.

The first result relates to our theoretical hypothesis and the empirical stalemate on folk-conceptions of free will. Our study shows that the folk-concept of free will is *apparent* hard incompatibilist, but that this appearance is caused by participants' misinterpreting probabilistic descriptions as entailing either 1) epiphenomenalism or 2) a conditional ability to do otherwise. At first

glance, we do find that laypeople ascribe free will significantly less frequently to agents in the probabilistic vignettes than to agents in the indeterministic vignette (stage 1). It would then seem that making the probabilistic aspect of indeterminism salient does indeed interfere with folk-conceptions of free will, such that they reveal themselves to be hard incompatibilists. However, this is only true when we ***do not*** control for understanding mistakes. Interestingly, a large number of participants misinterpreted the probabilistic vignettes to entail a conditional ability to do otherwise (31.5%). Here, we call this interpretation *conditionalizing*. In philosophy, indeterminism entails an *unconditional* ability to do otherwise whereas the conditional ability to do otherwise is related to determinism. Conditionalizing may then be seen as reflecting a *deterministic* (but not epiphenomenal) interpretation. Since a conditional reading of the probabilistic vignettes does not reflect an indeterministic interpretation as intended, the occurrence of conditionalizing was filtered as an understanding error (except in response to deterministic vignettes, where it reflects the philosophically accurate interpretation). When controlling for participants' understanding errors, the difference between free will ratings to indeterministic and probabilistic vignettes disappeared (stage 2). This reveals that people are only *apparent* hard incompatibilists. The apparent hard incompatibilist responses from stage 1 analysis thus do not reflect actual hard incompatibilist conceptions, but rather the fact that people don't as often interpret probabilistic descriptions as entailing an unconditional ability to do otherwise. When the probabilistic descriptions are correctly interpreted to include mental causality and an unconditional ability to do otherwise, any difference in free will ascriptions to agents in probabilistic and indeterministic vignettes disappears. Our results thus do not garner support for Kastrup's notion that we sometimes think of free will as non-random, at least regarding the kind of probabilistic randomness[17] that is a common focus in the free will literature (Van Inwagen 1983, 2017, Strawson 1986, Kane 2002, Mele 2006, Roskies 2006, 2010, Haggard 2010). Indeed, when looking at the fully-fledged free will profiles of participants who properly understood both vignettes they were presented with, only 6.5% (2 of 31) gave hard incompatibilist responses. Instead, our study shows that when laypeople understand the philosophically relevant concepts as philosophers intend them (a core assumption of the contrastive vignette method), they generally apply the incompatibilist concept of free will. This finding provides us with the first block of evidence to break the empirical stalemate about folk-conceptions of free will. Commonly, bypassing and intrusion effects have been used to problematize findings of previous studies,

[17] By talk of 'probabilistic randomness' we do not mean to allude to any particular alternative notion of randomness, but only to express openness to the idea that others may conceptualize randomness differently than the technical conception common in philosophy.

with bypassing problematizing incompatibilist findings and intrusion problematizing compatibilist findings. For instance, Nadelhoffer et al. show that compatibilist results, in particular those of Nahmias et al. 2005, can be explained by the presence of intrusion effects. But this just shows us that the results of the previous literature are driven by such effects, it does not show us which concept is at play when neither effect is present. For instance, it could be the case that although the compatibilist results of previous studies were driven by intrusion effects, laypeople would continue to give compatibilist answers when no intrusion or bypassing was present. The results presented here are the first to show which folk-concept of free will is in place behind the veil of common misunderstandings – i.e. when neither intrusion, bypassing, nor conditionalizing is present – and that it is incompatibilism. With this, we have the first block of evidence to show that *after* the mistake of bypassing is corrected, incompatibilism remains the dominant folk-conception. In 2005 Nahmias et al. noted about rollback scenarios that "such scenarios can also be varied to test whether indeterminism (e.g., one decision occurs half the time and another occurs half the time) increases or decreases participants' judgment of free will and moral responsibility." (Nahmias et al. 2005, p. 574). The present study is a first step towards testing whether such probabilistic considerations influence judgments of free will. However, we did not check for particular probability distributions such as a 50/50 chance. Further research is needed to determine if particular distributions influence participants' free will rating, and where the line lies – if there is one – between freedom-undermining randomness and non-random probabilities.

The second result is that even when vignettes include clear explicit descriptions of mental causation, about half of laypeople still interpret determinism to entail epiphenomenalism (49.4%). This demonstrates the strong prevalence and power of the bypassing effect. The strong presence of bypassing, despite including explicit descriptions of mental causation in the vignettes, has several possible explanations. First of all, it could reflect a problem with the design of the vignettes themselves. One thing of notice is the long length of the vignettes. Perhaps vignettes of this length are more cognitively demanding to read, causing participants to overlook the explicit description of mental causation and thereby drive the high presence of bypassing. However, a majority of participants understood the indeterministic vignettes perfectly fine, so it is unlikely that the effect can be explained entirely in terms of length and cognitive demand[18]. We also checked for a correlation between participants' level of education and the understanding scores of participants – if reading comprehension was a factor we should expect higher levels of education to correlate with a better understanding score,

[18] Using the flesh reading ease score our vignettes had a score of 57.9 (deterministic), 60 (indeterministic), and 51.5 (probabilistic). This corresponds to a difficulty level appropriate for 10-12th graders.

but we found no such correlation. Furthermore, where previous studies have shown effects of vignette length on free will ratings longer vignettes drive *compatibilist* responses (Nichols & Knobe 2007), and the bypassing effect drives *incompatibilist* responses. For these reasons, we find it unlikely that the strong presence of bypassing can be explained in terms of reading fluency, length of vignettes, or cognitive demand.

Defenders of compatibilism may point to another explanation for why the bypassing effect is so persistent: Perhaps it is because the folk-concept of free will is more about the epiphenomenalism/mental causation distinction instead of the determinism/indeterminism distinction. Nahmias et al. write that "if it takes a basic incompatibilist argument to make incompatibilism 'the intuitive view', then it seems that it is the incompatibilist who are talking the folk *into* incompatibilism" (Nahmias et al. 2005, p. 576). Similarly, they may argue that if it takes a basic learning process to make people think of free will as something related to indeterminism, then it is not compatibilists who talk the folk out of anything, but incompatibilists whose position must be learned, "taught - by professional philosophers, in philosophy books, and through philosophy courses" (Pink 2004, p. 43). Perhaps *metaphysics* is unintuitive and difficult, which would make it difficult to see how any concept with strict and nuanced ontological commitments could correspond to folk-conceptions of anything. Perhaps the ideas of indeterminism and determinism are themselves unintuitive and require a bit of learning to think (or intuit) about. It has, for instance, been shown that belief in free will is more closely related to beliefs in mind-body dualism than to belief in indeterminism (Wisniewski et al. 2019).

Although we agree that common-sense views of free will may be more closely tied to the ideas of substance dualism, mental causation, and epiphenomenalism, our results here show that it nevertheless remains the case that *when* people properly understand the concept of determinism, it clashes with their ordinary idea of free will. If determinism was intuitively irrelevant, it would indeed also be odd why philosophers have been so preoccupied with this part of metaphysics in the first place. As such, compatibilists should have to provide further reasons for why we should accept to call their concept *free* will when it is exactly what everyone else means by *unfree* will when they understand the implications of determinism.

Furthermore, the bypassing measure we apply here has been criticized for being overly sensitive. Björnsson and Pereboom propose that the bypassing effect may, in fact, be a 'throughpass' effect, where participants understand the mental states to be part of the causal structure but answer bypassing questions in the negative because they don't view these mental states as being the 'ultimate difference-maker' (Björnsson 2014, Björnsson and Pereboom 2016). Indeed, they conducted a study

that shows that most people who are apparent bypassers agree with a throughpass interpretation of deterministic vignettes. Those who appear to commit bypassing may then in fact not apply an epiphenomenalist reading after all. Here, we do not control for throughpass interpretation of the bypassing questions. However, since our vignettes explicitly include mention that beliefs, desires, and intentions affect human decision-making, and our bypass measure asks if these states have an effect on what a person does, our bypassing measure here doubles as a standard attention check. Furthermore, the bypassing/throughpass effect drives incompatibilist responses, so if our bypassing measure is overly sensitive and filters out non-epiphenomenal readings as epiphenomenal, we are likely filtering out real incompatibilist as 'apparent' incompatibilists. This would make our results that laypeople have incompatibilist conceptions more surprising. If many of our apparent bypassers are in fact genuine incompatibilists, then the real proportion of genuine incompatibilists is likely higher than what we find here.

In this experiment, we used exclusively non-moral vignettes. While it is common to find moral vignettes applied in the literature, we decided against it here because more emotionally salient vignettes have been shown results and result in more cases of intrusion (see Nichols & Knobes 2007, but for an exception see Nahmias et al. 2005, also Rose et al. 2015, and Nadelhoffer 2020). Keeping in line with the 'performance error' hypothesis (Nichols & Knobe 2007), which holds that strong affective reactions can bias and distort peoples' responses, and wanting to avoid invoking affective errors, we used exclusively non-moral vignettes. Proponents of the 'affective competence' hypothesis (Nichols & Knobe 2007), which on the contrary holds that the effects of strong affective reactions are not errors, might object that omitting moral vignettes our study 'stacks the deck' in favor of incompatibilist results. The lack of moral dimensions is a clear limitation of this study, and it would be interesting to see if the previously demonstrated discrepancies between responses to vignettes that evoke higher or lower emotional reactions remain when controlling for understanding errors, or if these previous findings are caused entirely by emotional salience being a source of intrusion. Further research is needed to investigate if this discrepancy remains when appropriate understanding controls are in place.

Regarding the 7-point Likert-type scale used here, our scale was only partially labeled (in contrast to fully-labeled scales commonly used in the literature). Though the effects of different scales and labeling approaches are a well-studied phenomena results are mixed (Andrews 1984, , Lau 2007, Weijters et al. 2010, Moors et al. 2014, Lewis 2019). Here, we ultimately decided to go with partial labeling which may reduce acquiescent response bias at the risk of resulting in more extreme response bias (Weijters et al. 2010). Another point that warrants discussion concerns the uneven recruitment of

participants for the two conditions. We recruited 66 participants for the determinism-indeterminism condition and 92 for the determinism-probability condition. While such a distribution is unorthodox, the purpose of this study was to check for responses to vignettes where *no* understanding errors were present. Based on a priori power testing, we needed at least 42 responses to each vignette. Because a high presence of bypassing and intrusion effect was expected, we ran understanding checks on the data while testing – but without doing any analysis of free will ratings – to ensure that we ultimately had enough responses *without* bypassing, intrusion, or conditionalizing effects for a meaningful analysis. These continued understanding checks revealed that to achieve the required sample size of at least 42 responses to probabilistic vignettes without any understanding errors, it was necessary to recruit more participants for the determinism-probability condition. This reflects the fact that understanding the probabilistic vignettes as intended was more difficult compared to the indeterministic vignette (see section 2.5).

Furthermore, it is possible that by presenting the two vignettes one after the other instead of both at once, the design does not successfully test intuitions. Perhaps participants answer intuitively to the first vignette, but more reflectively to the second vignette. However, if this was the case, we should expect to see an order effect. No order effects were found at stages 2 and 3. Similarly, we should expect a difference in free will ratings based on reaction times, with slower reaction times being indicative of more reflective i.e. less intuitive answers. However, no effects based on reaction times were found. Additionally, as noted previously (section 2.1), it is dubious whether previous research successfully taps into *intuitions* as such, which is also why we in this paper talk of folk-*conceptions* more generally instead of folk-intuitions.

Finally, as part of the understanding check participants were asked whether an agent 'could have done otherwise'. In the free will literature, this is a controversial phrase. Compatibilists frequently argue that 'could have done otherwise' is best understood in the conditional sense. Here, we control for this particular interpretation. The possible answers to the 'could have done otherwise'-question include the options 'only if her beliefs and desires had been different' and 'yes, even if her beliefs and desires were exactly the same'[19]. This explicitly gives participants the option to choose either the conditional or the unconditional interpretation of the wording.

[19] For the exact framing of the question and the other options for answering see supplementary notes 2.

2.7 Conclusion

Our study examined whether probabilistic randomness clashes with the common-sense notion of free will. We did not find any support for this hypothesis. Indeed, only a minority (6.5%) of participants who passed the understanding checks gave hard incompatibilist responses. However, we did discover a new kind of understanding error common for probabilistic vignettes. We call it conditionalizing: Interpreting indeterministic descriptions as including a conditional instead of an unconditional ability to do otherwise. On investigating the relationship between probabilistic randomness and common-sense notions of free will, our study is only a first step. There may be particular thresholds where specific probability distributions affect attributions of free will in indeterministic universes when such worlds are framed as the presence of probabilistic chance events and not simply as the absence of determinism. Nevertheless, we have shown that when controlling for the common understanding errors such as bypassing and intrusion effects, folk-conceptions of free will are incompatibilist. By controlling for these effects, we get a view of what the folk-conception of free will is when laypeople properly understand the philosophically relevant concepts, and that when this is the case laypeople are mostly incompatibilists. This comes as a challenge for those holding that compatibilism aligns with the common-sense notion of free will: They will have to show that there are cases where laypeople genuinely apply the compatibilist notion of free will, even when they understand deterministic descriptions properly.

3. Belief in Physicalism or Belief in Substance-Dualism: Free Will is Incompatibilist

Abstract

Belief in substance dualism – the idea that the mind and the mental is its own kind of substance, distinct from the physical – is a strong predictor of free will beliefs. Why? Here, we show that reduced belief in substance dualism appears to drive differences in free will *conceptions*. Specifically, we investigate if more physicalist beliefs – a particular kind of substance *monism* - drive incompatibilist and hard incompatibilist conceptions of free will. Using a within-subject design (N = 185), we find that people with more physicalist beliefs ascribe free will less often to agents in deterministic worlds, and thus appear to be more *incompatibilist* than their dualist counterparts. However, we show that this difference is not driven by a genuine difference in free will conceptions, but rather by differences in the (mis)understanding of deterministic descriptions: Physicalists interpret determinism as excluding mental causation. When this understanding error is controlled for, physicalists do not appear more incompatibilist – or hard incompatibilist – than dualists.

3.1 Introduction

Is the presence of probabilistic chance a problem for the physical mind? One of the most oft-discussed aspects of the free will debate, is whether free will is compatible with determinism or not, I.e. whether free will is compatible with the idea that at any instant only one future is compatible with the state of the universe at that instant and the laws of nature (Mele 2006). A more recent addition to this discussion is the question of whether the presence of probabilistic chance is compatible with free will. Born writes that "if you believe in perfect freedom you will get into difficulties again, because you cannot neglect the laws of statistics which are laws of nature" (Born 1949, p. 127), with Cassirer "it does not matter in this case whether one thinks of the causality governing nature in the form of strict dynamic laws, or of mere statistical laws. […] neither the one nor the other path leaves an opening for that sphere of freedom which ethics claims for itself" (Cassirer 1956, p 209), and Roskies writes that

"Randomness as the cause of action is as difficult to reconcile with a notion of freedom of the will […] as is determinism" (Roskies 2006).

However, it has been suggested that the focus on these compatibility questions is misguided and that the real threat to free will is instead reductionism and substance-monism: The idea that everything is fundamentally one substance (Nahmias 2006, Roskies 2006, Montague 2008, Cashmore 2010). Though there are many kinds of substance monism, such as neutral monism (Spinoza 1677), and idealism (Berkeley 1710), the most common form of monism in contemporary philosophy is *physicalism* (Bourget & Chalmer 2013). According to physicalism, everything is made of (or the result of) physical 'stuff', i.e. matter and the physical laws of nature (Tse 2011, Stoljar 2017). In contrast, substance-dualism is the idea that the mind and mental states consist of their own particular kind of substance; an immaterial, non-physical substance (Descartes 1641). Studies focusing on the relationship between free will and substance dualism show two seemingly inconsistent results. On the one hand, studies show that the ordinary concept of free will is not committed to any substance-dualist metaphysics (Monroe and Malle 2010, Monroe et al. 2014, Mele 2014, Nahmias et al. 2014, Vonasch et al. 2018,). On the other hand, belief in substance dualism better predicts belief in free will than (dis)belief in determinism does (Forstmann & Burgmer 2018, Wisniewski et al. 2019). These are odd findings. Why is belief in substance dualism a strong predictor of belief in free will if the ordinary free will concept is not committed to this kind of metaphysics? Furthermore, Wisniewski et al. found that belief in determinism *increased* free will belief amongst the more dualist participants, but decreased free will beliefs amongst monists in their western population[20].

Here, we investigate *why* belief in substance dualism might predict free will beliefs. It has been suggested that scientists, besides being physicalists, also tend to be *hard incompatibilists* (Shadlen & Roskies 2012). I.e. they think that not only is determinism a problem for free will, but that indeterminism is detrimental for our potential to have free will as well (Pereboom 2001, 2014a, Pereboom and Caruso 2002, Shadlen & Roskies 2012, Caruso 2012, 2019b). Here, we focus on the physicalist notion of substance monism and test both whether more physicalist beliefs drive 1) differences in free will *conceptions* and 2) differences in the understanding of the related *metaphysics* of determinism and indeterminism. We consider three different free will concepts: *Incompatibilism, compatibilism*, and *hard incompatibilism*. Incompatibilism is the notion that free will requires ontological indeterminism. Compatibilism on the contrary holds that no such metaphysical requirement is necessary

[20] The inverse relationship between determinism and free will beliefs was true for western participants only. For eastern participants, determinism was positively correlated with free will beliefs in both dualist and monist participants (Wisniewski et al. 2019).

for our potential to have free will. While most research on free will intuitions focuses exclusively on these two conceptions, we here include the position of hard incompatibilism. As previously noted, hard incompatibilism is the notion that free will is incompatible with both determinism and indeterminism.

Differences in free will conceptions can explain differences in free will beliefs. If more physicalist beliefs drive incompatibilist and hard incompatibilist free will conceptions, this would explain why belief in substance-dualism predicts belief in free will: Physicalists believe free will is impossible if determinism is true (incompatibilism) or even that free will is impossible regardless of whether determinism or indeterminism is true (hard incompatibilism), whereas dualists believe that free will is possible regardless of whether determinism or indeterminism is true (compatibilism). If this is the case, we should expect physicalists to be less inclined to believe in free will if they believe the world is deterministic or even if they believe the world is indeterministic.

In addition, physicalism may drive different interpretations of the concept of determinism. Studies show that people often misinterpret determinism (Nahmias & Murray 2010, Rose et al. 2015, Nadelhoffer et al. 2020). In particular, laypeople struggle to understand the philosophical notion of determinism in two different ways. The first interpretation challenge is known as the *bypassing effect*, whereby participants misunderstand determinism to exclude mental causation (i.e. to entail epiphenomenalism[21]) (Nahmias & Murray 2010). The second interpretation challenge is the *intrusion effect*, where participants intrude indeterminism into their interpretation of determinism (Rose et al. 2015, Nadelhoffer et al. 2020). These different understanding errors can also explain part of the results from Wisniewski et al., and it has indeed been shown that different beliefs about substance dualism and substance monism drive different interpretations of determinism (Forstmann & Burgmer 2018). Since intrusion drives high free will ratings and bypassing drives low free will ratings, if dualists have an intrusionist understanding of determinism and physicalists have a bypassing understanding, this could explain why determinism increases free will beliefs amongst dualists but reduce it amongst monists.

The general aim of this study is to contribute to the literature on folk-conceptions of free will by investigating if- and how differing beliefs in substance dualism or physicalism affect people's concept of free will. We hypothesize that 1) belief in physicalism leads to more incompatibilist and hard incompatibilist conceptions of free will, I.e. physicalist participants will ascribe free will less often to agents in deterministic worlds and to agents in worlds where indeterminism is described as the presence of probabilistic chance (in contrast to the mere absence of determinism), and 2) That belief in physicalism leads to more bypassing effects and that more dualist beliefs lead to more intrusion effects,

[21] Epiphenomenalism is the idea that mental states play no causal role in the world (Robinson 2019).

i.e. physicalists will tend to interpret deterministic descriptions as entailing epiphenomenalism whereas dualist participants will interpret deterministic descriptions as including indeterministic metaphysics.

3.2 Methods and materials

An online survey of 185 participants was conducted. Part of the participants was recruited via Prolific Academic (n = 96) and part of the participants was recruited via email (n = 89). Email recruitment was used to recruit scientists in order to gain a more physicalist sample. Due to a software error, gender data was only available for the participants recruited via prolific academic of whom 41 were female, 48 were male and 7 were unknown/preferred not to say. Participants' ages were in the range of 18-24 to above 65 years with a median range of 25-34 years. An a priori t-test using G*Power (Alpha = 0.05, power = 0.95, effect size = 0.8) was used to determine a minimum sample size of 84. Because the aim was to investigate conceptions of non-philosophers, 8 participants were excluded for having a background in philosophy (a Bachelor's degree or higher). Furthermore, 18 participants were excluded for failing an English language test presented at the beginning of the experiment, and 27 participants were excluded for failing an attention check (the 'eye-color' attention check, see later this section. See also supplementary notes 4). The final sample size after exclusions was then 132 participants. The experiment was built and run using *JavaScript* and lasted 12 minutes on average. Participants recruited from Prolific Academic were paid 1£. Participants recruited via mail received no payment for their participation. Ethical approval was granted by the ethics committee of the School of Advanced Study at the University of London (SASREC_1819-313C-R).

For this experiment, three different vignettes were designed describing three different worlds. The three kinds of worlds were 'deterministic', 'indeterministic', and 'probabilistic'. The difference between the indeterministic and probabilistic world is that the one we here call indeterministic describes determinism simply as the absence of determinism (i.e. 'if the same event happens at two different times, then the very same effect may or may not follow'), whereas the probabilistic description describes indeterminism as the presence of probabilistic chances. All descriptions include an explicit description of mental causation. One of the primary goals of the vignettes was to make the descriptions as easy to read and understand as possible, to reduce the presence of understanding errors compared to vignettes previously employed in the literature. Below are the three vignettes.

<u>Deterministic vignette (low-stake)</u>

"Please carefully read the text below. You will have to answer questions about it.

Imagine the following world: In this world, people are not puppets. What they do is caused by what they think, feel and decide.

In this world, everything follows **strictly** from cause-and-effect. Everything is strictly caused by what happened before. If the same event happens at two different times, then the very same effect follows.

One day, in this world, John decides to marry his girlfriend Laura. Like everything else, this decision was completely caused by what happened before it."

Indeterministic vignette (low-stake)

"Please carefully read the text below. You will have to answer questions about it.
Imagine the following world: In this world, people are not puppets. What they do is caused by what they think, feel and decide.

In this world, everything follows **partly** from cause-and-effect. Everything is partly caused by what happened before. If the same event happens at two different times, then the very same effect may or may not follow.

One day, in this world, John decides to marry his girlfriend Laura. Like everything else, this decision was partly caused by what happened before it."

Probabilistic vignette (low-stake)

"Please carefully read the text below. You will have to answer questions about it.
Imagine the following world: In this world, people are not puppets. What they do is caused by what they think, feel and decide.

In this world, everything happens as a matter of **probabilities**. Everything probabilistically results from what happened before. If the same event happens at two different times, then each time there is a certain probability, between 0 and 100%, that it is followed by a given effect.

One day, in this world, John decides to marry his girlfriend Laura. Like everything else, this decision probabilistically resulted from what happened before it."

Each participant read a consent form, a set of instructions and was assigned at random to one of two experimental conditions (Fig. 3, A). In both conditions, participants were presented with two vignettes. In condition 1 (determinism-indeterminism) participants were presented with the deterministic and the indeterministic vignette. In condition 2 (indeterminism-probability) participants were presented with one deterministic vignette and one probabilistic vignette. The order of vignettes was randomized.

After reading the first vignette, participants were asked "In the world you just read about, is it possible for a person to have free will?". Responses were recorded on a 7-point Likert scale ranging from "Yes, very certain" to "No, very certain" with the intermediate "Don't know" (see supplementary notes 5). After giving their rating, participants were asked to explain their answers. Following this, participants were asked two understanding questions, one to check for bypassing one to check for intrusion (see supplementary notes 4). Afterward, the procedure was repeated with the second vignette, before participants were asked demographic questions and a set of 5 dualism-questions adopted from the free will inventory (Nadelhoffer et al. 2014) designed to measure belief in substance dualism (see supplementary notes 6). Answers to the dualism-questions were recorded on a 6-point Likert-scale ranging from "strongly disagree" to "strongly agree".

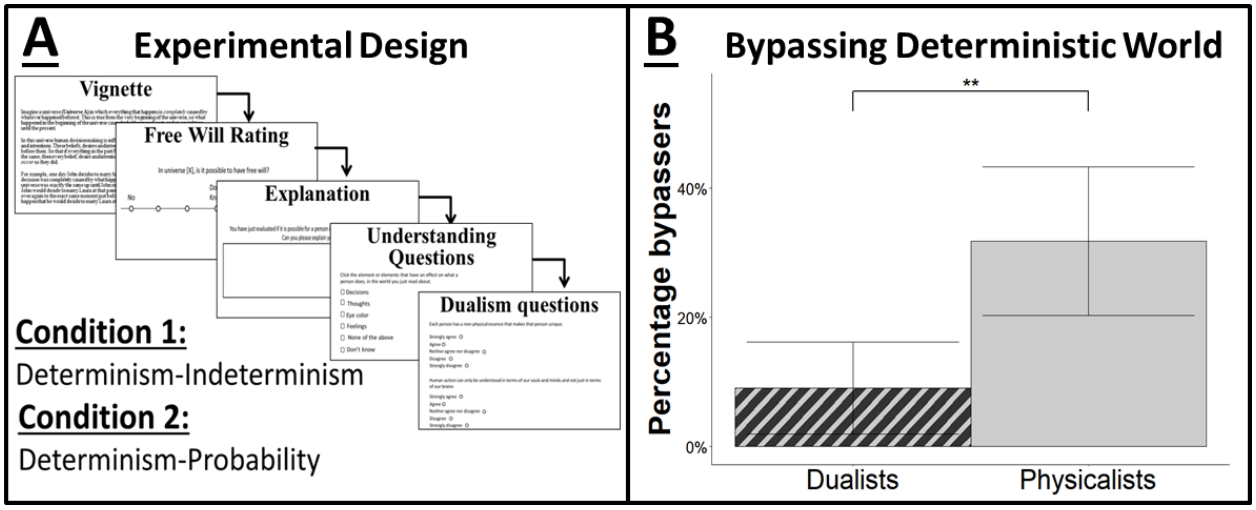


Fig 3. A. *Experimental design*. The procedure of the experiment was as follows: Participants are allocated either to the determinism-indeterminism condition or the determinism-probability condition. After that, they are presented with one vignette chosen at random and then asked to rate if it is possible to have free will in the described universe. They are then asked to explain their rating and asked understanding questions. Afterward, they are presented with the second vignette and the procedure is repeated. At the end of the experiment, participants are presented with the dualism questions. **B. *Bypassing responses to deterministic vignette.*** Show the percentage of responses by respectively dualists and physicalists that commit bypassing in response to the deterministic vignette. Error bars reflect confidence intervals.

The bypassing question asked participants to "click the element or elements that have an effect on what a person does, in the world you just read about". Clicking neither 'decisions', 'thoughts' nor

'feelings' reveals an epiphenomenalist interpretation and as such cases of bypassing. As part of the bypassing question, the option 'eye color' was included as a general attention check, as clicking 'eye-color' clearly revealed a lack of attention paid to the vignette. The intrusion question asked "In the world you just read about, if we go back in time to *just* before John made the decision to marry his girlfriend, could he have made a different decision?" Answering 'yes' in response to the deterministic vignette reveal an indeterministic reading of the vignette i.e. the presence of the intrusion effect.

For statistical analysis, Welch's t-test was used for comparing free will ratings. Chi-square was used for comparing distributions of fully-fledged free will profiles, and logistic regression was used for comparing the presence of bypassing and intrusion effects. When the data analyzed was a case of *complete separation*, running a logistic regression was not appropriate and so Fisher's exact test was used instead.

3.3 Results

65 participants were assigned the determinism-indeterminism condition and 67 participants were assigned to the determinism-probability condition. Replies to each of the 5 dualism questions were scored from 0 (strongly disagree) to 4 (strongly agree) to create a compound dualism-score (ranging from 0 to 20). The mean dualism-score was 11, the median was 11.5, and the standard error was 0.47. This shows generally more dualist beliefs in line with previous research (Nadelhoffer 2014, Wisniewski et al. 2019). We also found a significant difference in dualism-scores between scientists and non-scientists (Welch $t(4.8)$, $p < .001$), with scientists being more physicalist than their non-scientist counterparts. To check if more physicalist beliefs affect free will concepts we categorized participants who scored 11 or below as 'physicalists' and those with 12 or higher as 'dualist' (50/50 median split, $n = 132$).

Data was analyzed to examine three different aspects. First, the data was analyzed without controlling for either bypassing or intrusion effects, to examine the relationship between dualism-scores and free will ratings (stage 1). Secondly, we analyzed responses to the understanding scores, to check for bypassing and intrusion effects, and to analyze the relationship between dualism-scores and these understanding errors. Thirdly, using only responses where no bypassing or intrusion effects were detected, we again compared dualism scores and free will ratings (stage 2). At no point was any order-effect detected.

40% of participants ascribed free will to agents in deterministic worlds ($N = 132$), 87% ascribed free will to agents in indeterministic worlds ($n = 65$), and 67% ascribed free will to agents in probabilistic worlds ($n = 67$). When comparing free will ratings between physicalists and dualists, physicalists ascribed

free will less often to agents in deterministic worlds (Welch t(130) = 2.1, *p* = .03, Fig. 4, A). No difference was found between free will ratings of physicalists and dualists to indeterministic (Welch t(63) = -1.8, *p* = .06) and probabilistic (Welch t(63) = -0.1, *p* = .09) vignettes (Fig 4, B and C). When mapping out the fully-

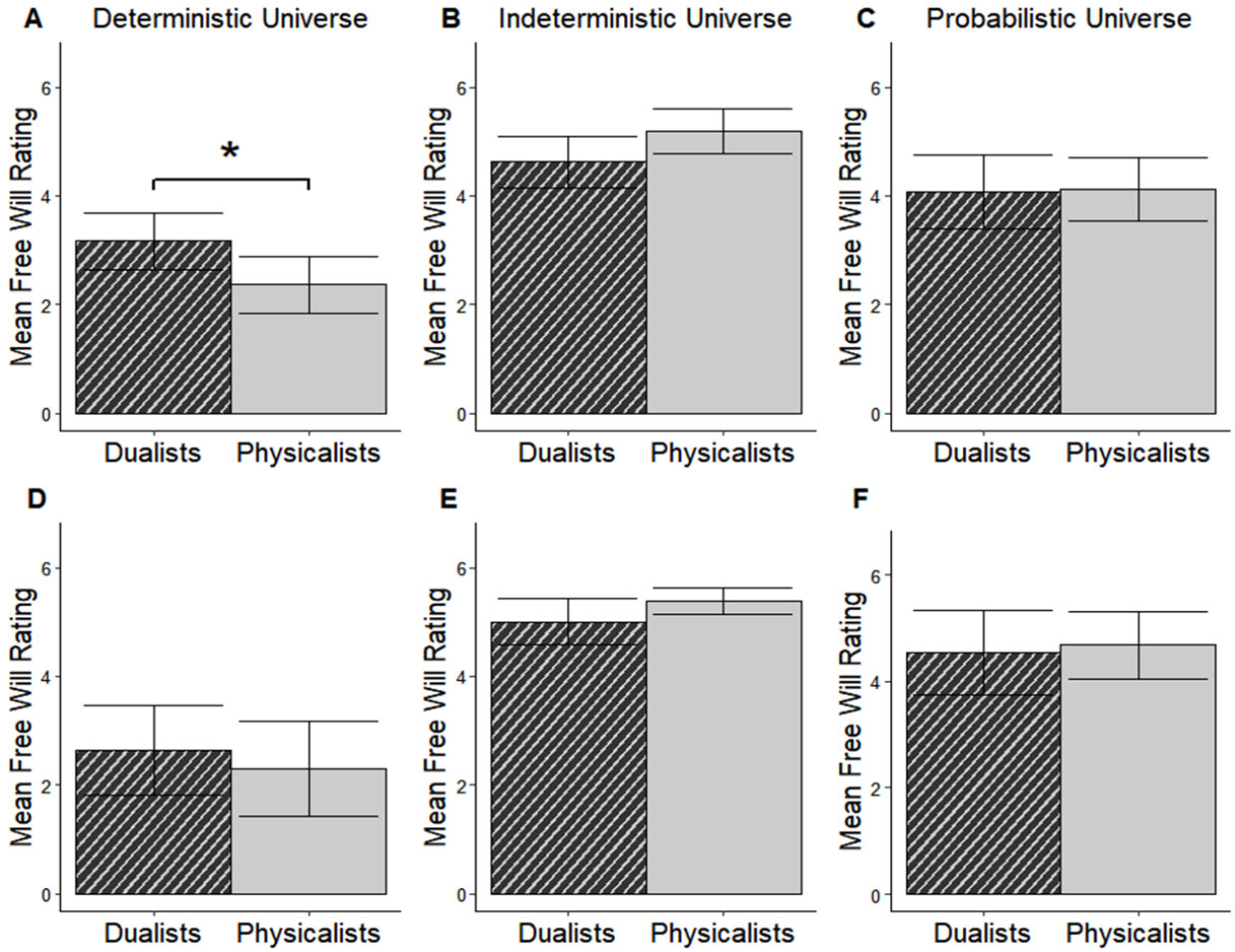


Fig 4. Free will ratings of dualists and physicalists. A-C shows ratings at stage 1 of analysis without correcting for bypassing and intrusion effects, D-F shows ratings at stage 2 of analysis after controlling for bypassing and intrusion effects. **A, D.** Deterministic. **B, E.** Indeterministic **C, F.** Probabilistic. Errors bars reflect confidence intervals.

fledged free will profiles of participants, 36% were compatibilist, 36% were incompatibilists, 9% were hard incompatibilists, 2% were exclusive determinists and for 17% we were unable to assess their precise profile because they had given a free will rating of "don't know" to one or more of the vignettes. When comparing the distribution of these free will profiles between physicalists and dualists a significant difference was found (x^2 p = .02, N = 132), with dualists being more compatibilists and physicalists being more incompatibilist.

When checking for bypassing and intrusion effects, 44% of responses to deterministic vignettes passed (n = 58), 84% of responses to indeterministic vignettes passed (n = 55) and 67% of responses to probabilistic vignettes passed (n = 45). When comparing differences in bypassing between physicalists

and dualists, there was a significant difference of bypassers in responses to the deterministic vignette (logistic regression, p = .002, Fig. 3, B). A marginally significant difference of bypassers was found in responses to indeterministic (Fisher's exact test, p = .49) vignettes and no significant difference was found in response to probabilistic (logistic regression, p = .78) vignettes. Also when comparing the presence of intrusion effects, no significant difference was found (logistic regression, p = .0530).

Of the responses with no bypassing or intrusion, 34% ascribed free will to agents in deterministic worlds (N = 58), 89% ascribed free will to agents in indeterministic worlds (N = 55), and 62% ascribed free will to agents in probabilistic worlds (N = 45). When comparing the free will ratings of physicalists and dualists to the different worlds, the difference between physicalists' and dualists' free will rating to deterministic vignettes disappeared (Welch t(55) = -0.6, p. = 55). There was also no difference between physicalist and dualist free will ratings to indeterministic (Welch t(39) = 1.6, p = .12) and probabilistic vignettes (Welch t(39) = 0.3, p = .79) (Fig. 4, D-F).

3.4 Discussion

Here, we find three primary results of interest. The first is that there is an *apparent* difference in the free will concepts between physicalists and dualists. Specifically, physicalists tend to ascribe free will less frequently to agents in deterministic worlds. Belief in physicalism thus appears to drive incompatibilist conceptions of free will, while dualist beliefs in contrast seem to drive more compatibilist conceptions of free will. One interpretation of this is that people who believe in physicalism are more likely to be incompatibilists and hard incompatibilists, in line with our first hypothesis. However, we did not find a significant difference in free will ascriptions between physicalists and dualists to neither indeterministic nor probabilistic worlds. As such, there is no support for the hypothesis that physicalists tend to be hard incompatibilists. It would then appear that physicalist beliefs drive more incompatibilist but not hard incompatibilist conceptions and that dualism drives more compatibilist conceptions. However, interpreting the data like this does not account for bypassing and intrusion effects.

The second primary result is that physicalist beliefs and dualist beliefs drive different kinds of understanding errors. Though the designed vignettes had a good readability score and should be fairly easy to read and understand (Flesch reading ease score = 66.5), less than half the participants understood determinism in the intended way. When comparing the number of bypassers between physicalists and dualists, we found that physicalists committed more bypassing in response to deterministic vignettes, but found no difference in the presence of bypassing in response to either indeterministic or probabilistic vignettes. This is in line with the explanation that the difference in free

will ascription between physicalists and dualists is driven by physicalists being more likely to interpret determinism to entail epiphenomenalism. Different interpretations of determinism based on belief in dualism and substance monism has been shown before (Paulhus & Carey 2011, Forstmann & Burgmer 2018), with monist beliefs driving belief in 'scientific determinism' and dualist belief driving belief in 'fatalistic determinism'. One particular criticism of the scales used in those studies is that the 'scientific determinism' subscale inaccurately describes action as caused by genes and upbringing instead of capturing the philosophical notion of determinism (Nadelhoffer et al. 2014). Here we expand upon this by showing that indeed physicalism, one particular kind of substance monism, drives epiphenomenal readings of determinism. This difference in understandings of determinism can explain why physicalists less often ascribe free will to agents in deterministic worlds. When comparing the frequency of intrusion effects between physicalists and dualists we did not find any significant difference. As such, we did not find support for the idea that dualist beliefs make people more likely to commit intrusion. However, the p-value reported here was very close to the threshold for significance (logistic regression, $p = .0530$). It is possible that, if an effect exists, it is too small for our sample size here to detect. Nevertheless, we here found support for the idea that physicalist beliefs drive epiphenomenal readings of determinism but no support for the hypothesis that dualists commit more intrusion than physicalists.

Finally, to examine if there was a real difference in free will conceptions, i.e. if any difference remained even after controlling for understanding errors, we compared free will ratings of dualists and physicalists using only responses where no understanding error was detected. On this analysis, however, no significant difference was found between physicalists' and dualists' free will ratings in response to any of the worlds. We thus failed to find support for the hypothesis that physicalists are *genuinely* more incompatibilist or hard incompatibilist than dualists. Any apparent difference in free will conception seems instead to be driven by physicalists being more likely to understand determinism to entail epiphenomenalism. When these differences in understanding errors are controlled for, the free will concept is stable across beliefs in physicalism and dualism.

It has been suggested that the fact that people 1) often reflect incompatibilist free will beliefs and 2) that belief in determinism positively predicts belief in free will, 3) reveals an inconsistency in people's beliefs (Wisniewski et al. 2019). This conclusion is pivotal on the idea that people interpret determinism in the philosophically relevant way. While we found no difference in intrusion between physicalists and dualists, there was a general very high presence of intrusion with more than a third of dualists (36%) committing intrusion. In other words, more than a third of substance dualists interpret determinism to include indeterministic metaphysics. Belief in this kind of indeterministic 'determinism'

is not inconsistent with the incompatibilist notion of free will. This throws doubt on whether previous findings really show an inconsistency in beliefs and further demonstrates the importance of controlling for participants' interpretation of determinism in this line of research. Whether the intrusion effect can fully account for the apparent inconsistency is an open question for future research to settle.

Previous research has shown that belief in substance dualism is not only intuitive but also a very persistent belief (Bloom 2004, Bering & Bjorklund 2004, Bering 2006, Hood et al. 2012, Hook & Farah 2013, Forstmann & Burgmer 2015). While differences in beliefs about substance dualism and physicalism don't drive differences in free will conceptions *per se*, our results might suggest a different reason why dualism beliefs are a strong predictor for free will beliefs: Perhaps belief in substance dualism simply tracks belief in *mental causation*, as physicalists are more likely to interpret surrounding beliefs (such as determinism) as entailing epiphenomenalism.

3.5 Conclusion

Here, we investigated the relationship between belief in physicalism or substance dualism and free will concepts. We found that physicalists and dualists have *apparent* differences in their free will conceptions, with dualists being more compatibilist and physicalists being more incompatibilist, but with no indication that physicalist beliefs lead to more hard incompatibilist conceptions of free will, contrary to Shadlen's and Roskies' claim that scientists tend to be both physicalists and hard incompatibilists (Shadlen & Roskies 2012). We also found support for the hypothesis that physicalist beliefs lead to more bypassing errors when interpreting deterministic descriptions, but found no support for the hypothesis that dualist beliefs lead to more intrusion. Furthermore, when the understanding errors of bypassing and intrusion were controlled for, there did not appear to be any difference in free will concepts between physicalists and dualists, revealing that any such difference was merely apparent and driven by differences in understandings of determinism.

4. Epistemic Virtues and Underdetermination: When Should We Be Indeterminists?

Abstract

In both philosophy and science, there is a common-place assumption that ontological determinism, the idea that everything follows with strict necessity from its causes, is real. Here, I argue that our main epistemological approaches justify the opposite conclusion. The main argument comes from the approach of epistemic virtues. I claim that it is only by accepting ontological indeterminism – the idea that determinism is false - that one can satisfy the epistemic virtue of being genuinely self-critical. From the position of virtue epistemology, this provides a sufficient reason for accepting indeterminism. I further consider two alternative epistemological approaches, direct scientific realism, and epistemic pragmatism. I argue that the first should lead us to accept ontological indeterminism as the probabilities present in scientific models should be taken as a direct sign of indeterminism. The second approach, by applying the maxim of explanatory use as truth, also gives us reason to accept ontological indeterminism.

4.1 Introduction

"*The assumption of an absolute determinism is the essential foundation of every scientific inquiry.*"
– Max Planck (Planck 1958, as translated by Heilbron 1986).

Nature is often thought to follow laws of strict necessity. This means that at every given moment only one specific future is compatible with the state of the universe and its laws. In other words, nature is thought to be deterministic. Here, I show that despite the common-place belief in determinism we actually have better reasons to accept ontological indeterminism; the idea that determinism is false and that, at any given moment, several possible futures are compatible with the state of the universe and its laws. I take an epistemology first approach, to show that regardless of which epistemology we adopt,

we find sufficient reasons to accept ontological indeterminism: The idea that the universe really *is* indeterministic, and does not merely appear to be so.

The line of reasoning considers three epistemological approaches: Direct scientific realism, virtue epistemology, and pragmatism. The main argument is rooted in virtue epistemology. Specifically, I claim that with the information available to us today it is impossible to be genuinely self-critical without accepting ontological indeterminism. This argument has implications for stochastic[22] models in general, but for the purpose of this paper, it focuses on models of the brain and mind. There are two reasons for this: First, ontologically indeterministic interpretations are still considered problematic in this area (see Gessell 2017). Secondly, interpretations of models of the brain and mind have far-reaching implications for other areas of philosophy. In particular for philosophy of action where the truth or falsity of indeterminism plays a key role in the free will debate and for our understanding of what it means to be able to 'do otherwise'.

Stochastic models of neural processes, cognition, and behavior are widespread. Explanations of these phenomena are written in the language of probabilities. From lower-level mechanisms of single neurons and their interactions[23], to the higher-level neural computations such as the computation of subjective value[24], and to the behavioral level of human decision-making[25], probabilities are part of the explanations: "Neuroscience requires probabilities" (Gessell 2017, p. 1220). But how should we interpret these probabilities? Are they a consequence of our inability to represent the world with perfect accuracy or is it an ontologically real aspect of the world our models represent?

Whether the world is fundamentally deterministic or genuinely indeterministic is a central question in the philosophy of probabilities. The question concerns how to *interpret probabilities*. The *propensity theoretical* interpretation of probabilities takes probabilities to reflect physical *dispositions* or *tendencies* (Pierce 1931, Popper 1959, Hacking 1965, Mellor 1971): Something 'physically real' "comparable to Newtonian forces" (Popper 1959, p. 27). In this view, chance is a *property* of the world, not merely an epistemically useful construct. In other words, propensity theory is a fundamentally *indeterministic* interpretation of probabilities (Gillies 1973, Kyberg 1974). As Mellor puts it, "if propensities are ever displayed, determinism is false" (Mellor 1971, p. 151). On the other hand, the *subjective* interpretation of probabilities considers probabilities as reflecting *degrees of belief* or

[22] By stochastic I mean a random process or event, i.e. a process which evolves according to probabilistic laws. (Werndl 2009).
[23] See for instance Dean 1981, Tolhurst et al. 1981, Britten et al. 1996, Rieke et al. 1997, Shadlen & Newsome 1998, Auger & Marty 2000, Alvarez et al. 2002, Platkiewicz & Brette 2010, Mohan et al. 2011, Wang et al. 2016.
[24] See Rieke & Baylor 1998, Del Cul et al. 2007, Kurtz-David et al. 2019.
[25] See Neuringer 1986, McFadden 2001, Schurger et al. 2012, Agranov & Ortoleva 2017.

credence (Talbott 2016). With the subjective interpretation "Probability is interpreted […] epistemically" (Hartmann & Sprenger 2010, p. 2). Or with de Finetti's words "The fundamental point of the subjectivist conception is that the notion of probability does not refer to something which is a property of the 'outside world'" (de Finetti 1974, p. 265). Probabilities are then something *subjective*: They reflect how likely a given outcome seems from our limited epistemological position of incomplete information. This is a position of *epistemic indeterminism* but *ontological determinism*[26].

In the models from cognitive science, quantum probabilities are employed increasingly more often and are often understood to imply something like propensities (Pothos et al. 2013). And yet, it remains a common-place assumption that all behavior and its underlying phenomena are, despite the success of probabilistic explanations, fundamentally deterministic. Renowned neuroscientist Patrick Haggard says that "As a Neuroscientist, you've got to be a determinist" (Haggard 2010, para. 15), perfectly in line with the Planckian maxim quoted at the beginning of this paper. It is a common conclusion that our current neuroscientific insights only suffice for a justification of epistemic, but not ontological, indeterminism[27] (Gessell 2017). Similarly, we often find the assumption of determinism amongst philosophers. A plethora of work discusses determinism and free will, but it is rare to find reasons for why we should be determinists in the first place (Fischer & Ravizza 1998, List 2014). These discussions are instead underlined by an implicit link between determinism and naturalism: An association so strong that the terms are sometimes used synonymously (see Dennett 2013, but for an exception see McKenna & Pereboom 2016).

Contrary to this picture and the belief in determinism, I show that we have better reasons to accept ontological indeterminism. That is to say, we should interpret the probabilities in our cognitive science models as signs of genuine indeterminism. Before proceeding, further two common objections should be put aside. These objections inevitably appear whenever arguments for ontological indeterminism of the brain and mind are presented. One relates to quantum indeterminism, the other to meta-metaphysics.

[26] Some Bayesians consider propensity theory a variety of Bayesianism (see Good 1983). However, Bayesianism is everywhere called the subjectivist interpretation (Kaye 1988, Hartmann & Sprenger 2010, Talbott 2016) and Popper specifically puts forth propensity theory in contrast to the subjectivist interpretations (Popper 1959). Conflating propensity theory and Bayesianism is then clearly a misclassification.

[27] Gessell speaks of *soft indeterminism* and *hard indeterminism* with the former being the position of epistemic indeterminism but ontological determinism that he advocates and the latter being the position of genuine ontological indeterminism. I will avoid this terminology of *soft* and *hard* indeterminism to avoid confusion with the terminology in the free will debate where soft and hard *determinism* are well-established positions. In the context of free will, these prefixes *soft* and *hard* denote something quite different from what Gessell has in mind. For the sake of clarity, I will therefore continue the use of the terminology *epistemic* and *ontological* indeterminism.

The first objection appears any time indeterminism of the brain is discussed along with quantum mechanics. The objection can be put like this: 'Whereas quantum-level phenomena may be ontologically indeterministic, this does not mean that macro-level phenomena like the ones we are examining in the cognitive sciences share this property. The peculiarities of sub-atomic particles have no direct relevance for the mechanisms of the brain.' (see Weber 2005 for an example of this kind of objection). Indeed we should be careful not to assume too much similarity between the parts of a thing and the whole of it. But the point I make is ***not*** an attempt to bridge quantum-level indeterminism to the macro-level operations of the cognitive sciences. What I propose is that we have sufficient reasons to accept macro-level ontological indeterminism, even without such a bridge[28]. The point I make by turning our attention to quantum mechanics is a note against deterministic intuitions and assumptions. As such, it is a different point than the kind this objection is (or should be) raised against.

The second kind of objection that one encounters when arguing in favor of ontological indeterminism, is that despite the reasons presented in favor of indeterminism, it remains *possible* that determinism is the case. Certainly, providing strong reasons in favor of one position over another often leaves the alternative possible. Yet when it comes to metaphysics, people seem to hold that unless the position they find intuitive can be shown to be decidedly *impossible*, then its mere possibility is sufficient reason to accept it. The argument I will present will also ***not*** be an argument from necessity. And when all is said and done it will remain quite possible that determinism is the case. However, anyone who hopes for an argument with the strength to deem either position impossible will surely be disappointed by any discussion on the matter whatsoever. All we can hope for is that when we weigh the reasons for and against the different possibilities, the weighing should turn out uneven so that one position has at least slightly better reasons to be accepted than its alternative. At the end of the present discussion, ontological determinism remains a *possibility*. But if that is all one has left to say in defense of the determinist position, then our judgment should be tipped in favor of ontological indeterminism.

With these clarifications made, the paper proceeds as follows. I will first give reasons for why we should not let our intuitions, but epistemological principles, guide our beliefs. Afterwards, I show that regardless of which epistemology we choose, direct scientific realism, virtue epistemology, or pragmatism, we have better reasons to accept ontological indeterminism. Potential objections will be discussed along the way.

[28] Nevertheless, it should be noted that wave-particle duality has also been demonstrated at the level of molecules (Arndt et al. 1999, Gerlich et al. 2011, Juffman et al. 2012), so it may not be long before one would have to insist that molecular interactions are of no relevance for the mechanisms of the brain, to keep the indeterminism out of it.

4.2 Quantum physics and the old-school determinists: A cautionary tale

In the early days of quantum mechanics, physics was confronted with the question of whether the world was ontologically indeterministic or not. Classical mechanics, which had been successful until then, corroborated a deterministic picture of the natural world. But the newly discovered quantum phenomena did not lend themselves easily to deterministic explanations. Opinions were split on how to interpret this probabilistic aspect of quantum mechanics.

Scientists such as Einstein and Schrödinger interpreted quantum phenomena in light of *classical* mechanics. They held that the world was genuinely deterministic and that quantum mechanics with its probabilistic explanations was *incomplete* (Schrödinger 1935, Einstein et al. 1935, Einstein 1936). The seemingly indeterminate nature of quantum phenomena was hypothesized as being caused by so-called *local hidden variables* – the idea being that the stochasticity of quantum mechanics was an epistemic consequence of our ignorance about the details of these hidden variables: A consequence of our epistemic limitations. On this view, the physical laws and quantum phenomena were thought to be ontologically deterministic. Our quantum mechanical descriptions would simply be incomplete descriptions due to a lack of information about these hidden variables. In opposition to this, other thinkers such as Heisenberg, Bohr, and Born accepted the *prima facie* indeterministic picture. They took the probabilistic descriptions of quantum mechanics to be a sign of ontological indeterminism (Born 1949). And so, the world of physics was divided on how to interpret the probabilities in their models. Were they merely a consequence of our epistemic limitations (the byproduct of an incomplete picture), or were they a sign of ontological indeterminism (a genuine chance-property in the laws of nature)? The question was much debated, with the most famous discussion on the matter perhaps being the Bohr-Einstein debates ranging from 1927 to 1935 (recounted in Schilpp 1959).

This difference between these two schools of thought about how to interpret quantum mechanics, the difference between deterministic *hidden-variables interpretations* and the indeterministic *Copenhagen interpretation* of quantum mechanics, is perhaps best illustrated in Einstein's now-famous picture from his letter to Max Born.

> "You believe in the God who plays dice, and I in complete law and order in a world which objectively exists, and which I, in a wildly speculative way, am trying to capture. I firmly believe, but I hope that someone will discover a more realistic way, or rather a more tangible basis than it has been my lot to find. Even the great initial success of the quantum theory does not make me believe in the fundamental

dice-game […]. No doubt the day will come when we will see whose instinctive attitude was the correct one." – Einstein, letter to Max Born, 1944 (Einstein et al. 1971, p. 149).

Here we then see the difference between determinism and indeterminism illustrated as follows: on the one hand, we have determinism as a world of 'complete law and order' where everything follows with strict necessity from what went before it. And on the other hand, indeterminism is described as an ontological 'dice-game', a genuine chance-element in the inner workings of the world. Einstein readily admits that his belief in determinism rests on an '*instinctive attitude*'; ultimately a matter of intuition. Today it seems that many still share this intuition and instinct, on which they base their beliefs about the ontological question.

Optimistically, Einstein asserted that the day would come when *we can see* which instinctive attitude is the correct one. In 1964, a short nine years after Einstein's death, the first step towards such a sight was made. John Stewart Bell formulated *Bell's theorem* (Bell 1964). The theorem reveals something rather counterintuitive: The deterministic local-hidden-variables kinds of theories entail different empirical outcomes than indeterministic Copenhagen-like alternatives. Eight years after the discovery of Bell's theorem we got a glimpse of which instinctive attitude was the correct one. The first Bell-type experiment was conducted (Freedman & Clauser 1972). To Bell's own surprise the outcome of this experiment fell in favor of ontological indeterminism. It provided evidence against local-hidden-variable interpretations of quantum mechanics. The unexpectedness of this result is well captured by Bell's comments on the outcome:

"For me, it is so reasonable to assume that the photons in those experiments carry with them programs, which have been correlated in advance, telling them how to behave. This is so rational that I think that when Einstein saw that, and the others refused to see it, *he* was the rational man. The other people, although history has justified them, were burying their heads in the sand. [...] So for me, it is a pity that Einstein's idea doesn't work. The reasonable thing just doesn't work." – John Stewart Bell, as quoted by Bernstein (Bernstein 1991, p. 84).

Since 1972 many different Bell-type experiments have been carried out, accounting for and testing various so-called 'loopholes' of the original experiment. To date, every single one of these tests has corroborated the original findings. The time may then have come where we can see whose

instinctive attitude was the correct one, and that it is Einstein who was wrong[29]. These days, quantum mechanics are often taken to reveal genuine "randomness *per se*" (Lecoutre et al. 2005, p. 22). If there is anything to learn from the peculiarity of the world, it is that we should be careful not to let our intuitions carry too much weight with regards to our beliefs. We must be careful not to let our instinctive attitudes have too much influence on what we believe about the machinery of the world.

But even with the evidence for ontological indeterminism on a quantum level, there is no guarantee that quantum indeterminism should be amplifiable to the levels of operations of other sciences. Perhaps the indeterminism is 'washed out' on the macro-level of personhood, mind, and action. Perhaps the world is 'layered' in different layers of indeterminism and determinism (Bohm 1957, Butterfield 2012). Nevertheless, the time is ripe for the cognitive sciences to also have this discussion about how to interpret probabilities. As noted, my aim is not to lay the bricks of a bridge stretching from the indeterminism of quantum physics to biology. But we now have reason to be skeptical of *instinctive* attitudes and intuitions, which paves the way for arguments based on an *epistemic* attitude and reasons. In the following, we will first consider the epistemology of direct scientific realism. Examining this approach and its shortcomings will in a natural way lead us to the second argument.

4.3 The first argument: Direct scientific realism and the naïve empiricist

The first attitude towards probabilistic models we will consider is direct scientific realism. An example of this line of thinking can be found in Fine's 'Naturalistic Ontological Attitude' (Fine 1986). The naturalistic ontological attitude states that "Our best guide to ontological questions in any domain is given by our best scientific theories of that domain" (List 2014, p. 12). Having a clear picture of this approach will be relevant for what comes later. Accepting this attitude without further justification and taking empirical models at face-value is what characterizes the position of the naïve empiricist[30].

The reasoning of this approach may proceed in the following way. Our naïve empiricist, let's call him Charlie, runs into the following consideration: Since probabilistic models include an indeterminate element and deterministic models do not, it seems that probabilistic models fit better with an indeterministic world than with a deterministic one. This is so because in an indeterministic world the

[29] It is worth noting, that while the local-hidden-variable variety of deterministic interpretations of quantum mechanics have fallen out of favor, other deterministic interpretations have been brought forth: So-called non-local-hidden-variable theories such as the Broglie-Bohm theory. However, even non-local-hidden-variable theories have more recently been problematized by experimental outcomes (Gröblacher et al. 2007, Paterek et al. 2007).
[30] Another example of this kind of argument can be found in Brendon and Carson in the context of evolutionary theory (Brandon & Carson 1996).

use of probabilities, i.e. the inclusion of an *indeterminate element*, is a perfect fit - contra the exclusion of such an element, as is the case with deterministic models (This should not be confused with the idea that the assigned probabilities – the actual quantification – therefore necessarily are a perfect fit). In light of this consideration, Charlie concludes that the success of probabilistic models (E1) is more likely in a genuinely indeterministic world (H1) than in a deterministic world (H0). He accepts this without further justification (exemplifying the naïveté of his position).

Formally his view on the likelihood of successful probabilistic models can be expressed as $p(E|H1) > p(E|H0)$. In light of empirical best-explanation models that are probabilistic, Charlie then updates his beliefs in such a way that he concludes ontological indeterminism to be the most likely possibility.

Formally this can be expressed as follows:

1. Probabilistic best-explanation empirical models (E) are more probable given ontological indeterminism (H1) than with ontological determinism (H0).
2. As open- and fair-minded reasoners, we should continuously update our beliefs in light of available information. This includes our metaphysical beliefs.
3. As a consequence of the evidential asymmetry (1), we should then conclude that the world is indeterministic (H1) in light of best-explanation probabilistic models (E).

This is the argument from scientific realism, and "the force behind it is a firm realist commitment, which underwrites the sharing of properties between the model and its system." (Gessell 2017, p. 1208). In Gessell's mouth, models are our scientific models and the system is the (part of the) world that the models represent. The argument may fall flat to many because of the naïveté of the first premise. In the following, we will consider an objection that attacks this premise. Doing so will clarify what we must keep in mind to accept the objection and reject the reasoning of the naïve empiricist. This, in turn, will be important for the argument to follow.

4.4 Objection to the first argument: Empirical evidence can tell us nothing of metaphysics

This objection attacks the first premise of the previous argument. The objection rejects the idea that best-explanation probabilistic models are more likely given ontological indeterminism than with ontological determinism.

The critic objects to Charlie's assumption that probabilistic best-explanation models are more likely in an ontologically indeterministic world than in an ontologically deterministic one. Instead, the critic claims, best-explanation probabilistic models are ***exactly*** what we should expect in an ontologically **deterministic** world *when we are in a position of incomplete information*. And since we are in a position of incomplete information, best-explanation probabilistic models are no more sign of ontological indeterminism than determinism. In other words, it is ***not*** any more likely that we will have best-explanation probabilistic models in an ontologically indeterministic world (H1) than in an ontologically deterministic one (H0).

Formalized, the objection goes as follows:

1. From a position of incomplete information, ontological determinism entails best-explanation probabilistic models.
2. We are in a position of incomplete information.
3. Best-explanation probabilistic models (E), is exactly as likely in a deterministic world as in an indeterministic world ($E|H0 = E|H1$).
4. Since the evidence E is equally likely in a deterministic world as in an indeterministic world, it can tell us nothing about whether the world is deterministic or indeterministic.

This is essentially the objection of empirical underdetermination (Best known as the Duhem-Quine thesis (Quine 1951, 1955, Duhem 1954, Suppes 1993, Stanford 2017)). In light of this objection, it seems that empirical insights can never inform us about this part of metaphysics. Indeed, the work of Werndl has shown that there are cases of empirical equivalence between deterministic and indeterministic models (Werndl 2009, 2011). In other words, there are cases where two models, one deterministic and the other indeterministic, describe a given system equally well – they make the same predictions and are interchangeable. Perhaps our neuroscientific phenomena are such cases, and we just don't have these equally well-working deterministic models yet.

At this junction, one might argue that since empirical insight can tell us nothing about metaphysics, we should in fact *not conclude anything at all* about this matter. That it is a folly to try and answer such questions with reason, and that we should abandon the project altogether. Putting aside this abandonment of metaphysics for a moment, let us continue taking the metaphysical question seriously. If we then accept the objection and abandon the approach of direct scientific realism, which epistemological alternatives remain? One alternative is the approach of virtue epistemology. This brings us to the second argument in favor of ontological indeterminism: The argument from epistemic attitude.

4.5 The second argument: Self-criticism and the argument from epistemic attitude

When practice can tell us nothing, we must consult the world of principles. The argument from epistemic attitude rests on the idea that *being self-critical* is an epistemic virtue. The argument shows that in light of current information, it is ***only*** by being indeterminists that we can satisfy the virtue of having a genuinely critical attitude towards our convictions.

In face of empirical underdetermination, Duhem posited that we have an intuitive reasoning ability, a 'good sense', that allows us to judge among competing theories, despite several theories being logically compatible with the available evidence (Duhem 1954). Recently, this part of Duhem's philosophy has been analyzed in terms of responsibilist virtue epistemology[31] (Stump 2007, Kidd 2011, for a criticism see Ivanova 2010). Responsibilists think of intellectual virtues as good *character traits* (Code 1987, Zagzebski 1996). The focus here is then the virtues *of the epistemic agent* (*practical* virtues), and not the virtues *theories* or particular kinds of explanation (*theoretical* virtues) [32]. The argument from epistemic attitude aims to draw a line for when we should change our mind in special cases of empirical underdetermination. As such it may be considered an elaboration of at least part of what constitutes Duhemian 'good sense'.

The argument starts with a clarification of what it is to be critical:

1. To be critical entails being self-critical.
2. Being self-critical entails taking one's own fallibility seriously.

[31] In contrast to *reliabilist* virtue epistemology (Sosa 1980, Goldman 1992, Greco 1992, 2003).
[32] For a discussion of the latter see for instance Weiskopf 2017, Keas 2018, and Taylor 2019.

3. Taking one's own fallibility seriously entails believing *in a way** so that we could learn (if not in practice, then in principle) ***that*** we are wrong ***if*** we are wrong (*the principle of a critical attitude*).
4. Believing *in a way* so we could learn that we are wrong if we are wrong, entails choosing beliefs for which we can imagine and think of a way for reality to reveal to us *that* we are wrong *if* we are wrong.

*By *way of believing* I mean the way in which we shape our convictions: How we move from merely considering an option to accepting and believing it, and how we move from merely making further considerations to updating our beliefs.

From this definition of being self-critical, i.e. as believing in a way so that we could learn that we are wrong if we are wrong, it may seem unclear how it differs from open-mindedness. Although definitions of open-mindedness vary (Hare 1979, Adler 2004, Riggs 2010), Riggs identifies it as "to be aware of one's fallibility as a believer, and to be willing to acknowledge the possibility that anytime one believes something, *it is possible to be wrong*." (Riggs 2010, p. 180). Open-mindedness is then simply the *acknowledgement* of one's fallibility, whereas being self-critical is to take one's fallibility seriously by believing in a specific *way*: Namely, in accordance with the principle of a critical attitude. To be self-critical requires that one is open-minded, but it is possible to be open-minded without thereby being self-critical. An example of the latter is the stereotypical mentality of conspiracy theorists, who accept that their beliefs may be wrong but for whom nothing could ever, even in principle, change their mind. Someone who is open-minded could agree with the sentence "It is possible I am wrong, but I can't think of anything that could change my mind about it", whereas someone who is self-critical could not[33].

As stated, to be self-critical entails believing in a way so that we could learn, at least in principle, that we are wrong *if* we are wrong. I.e. believing in a way so that, in a world where our current belief is false, there is a way (at least in principle) for us to learn *that* we are wrong. The argument from epistemic attitude then confronts us with the following question:

How could we, in principle if not in practice, learn that we are wrong if we are wrong?

[33] I take the three primary epistemic virtues to be *open-mindedness*, *fair-mindedness,* and *self-criticism*, understood in the following ways. To be open-minded is to acknowledge one's fallibility. To be fair-minded is to use the same principles and standards of evaluation for each of a set of competing ideas. To be self-critical is to believe in a way so that it is possible, at least in principle, to learn that we are wrong if we are in fact wrong. Though there is often overlap between these three virtues, it is possible to be both open-minded and fair-minded without necessarily being self-critical. A full elaboration of these three virtues is beyond the scope of this paper, however.

The indeterminist can answer as follows:

"*If* we are wrong, then, as we approximate complete information, we will acquire best-explanation deterministic models whereby we would learn *that* we are wrong."[34]

The determinists may be tempted to mirror this answer, and reply as well:

"*If* we are wrong, then, as we approximate complete information, we would acquire best-explanation *indeterministic* models, whereby we would learn *that* we are wrong."

The problem with the determinists' answer is that what they describe, as the thing that could teach us that we are wrong if we are wrong, is *exactly* what we find today. We *are* approximating complete information: We are continuously learning about the world. And as we continue to learn about the world, the best-explanation models we acquire *are* indeterministic – at least for the time being.

By mirroring the indeterminist's answer, the determinists re-introduce the asymmetry in evidential support that is the first premise of our naïve empiricist. They re-introduce the idea that best-explanation indeterministic models serve as evidential support against ontological determinism. This is the asymmetry that was *rejected* earlier. And so, the first argument resumes its validity, but this time on critical grounds.

Formally the argument from epistemic attitude goes as follows:

1. As fallible creatures, we should apply a critical epistemological attitude when shaping our beliefs.
2. Having a critical epistemological attitude entails believing in a way so that we could learn (if not in practice, then in principle) *that* we are wrong *if* we are wrong.

[34] It is important to note that the answer does not depend on the *possibility* of actually acquiring complete information. How close to complete information must we be to find best-explanation deterministic models in a deterministic world? If we have complete information (100%), then we will surely have such models in a deterministic world. But perhaps deterministic models overwhelmingly become best explanation models much sooner than that. Perhaps deterministic models will out-compete indeterministic models when we have acquired 70% of all information, or2.5%, or 0.000001%. It is impossible to tell. The important point is that such a breaking point where deterministic models become best-explanation models (or at least equally-well-explaining models) must *necessarily* exist in a deterministic world. And this breaking point could lie long before we reach complete information. As such, the impossibility of actually acquiring complete information does not invalidate the indeterminist's answer.

3. In light of current best-explanation probabilistic models, only **indeterminists** can learn (in principle) that they are wrong if they are wrong.
4. In light of current best-explanation probabilistic models, we should be indeterminists.

The question from a critical epistemic attitude remains as a standing challenge to determinists: To answer the following without either being uncritical and dogmatic, and without introducing the kind of evidential asymmetry that justifies indeterminism on empirical grounds.

If we accept determinism, how could we, in principle if not in practice, learn that we are wrong if we are wrong?

Without a determinist answer to this question, the reasonable conclusion is that *only* by being indeterminists could we ever learn (in principle if not in practice) that we are wrong if we are wrong. And so it is only by accepting indeterminism, in light of our current best-explanation models of the world, that we apply a genuinely critical epistemic attitude. Determinism on the contrary becomes a position of dogma.

The argument from epistemic attitude is particularly relevant for those who find that instrumental noise could justify a belief in determinism. For if we take the stochasticity of our models to be signs of instrumental noise, we put ourselves in the kind of position where we could never learn that we are wrong *if* we are wrong. In an ontologically indeterminist world, we would forever misinterpret the real indeterminism as instrumental shortcomings. To be genuinely self-critical, we must find another justification for a belief in determinism, than the fact that the stochasticity of our models *could* be the result of instrumental noise.

The metaphysician may again object, however, that there is no way we could learn even *in principle* that we are wrong if we are wrong. The proper conclusion, following from the second premise of the argument, is then that we shouldn't believe either position: That we can draw no conclusion at all. This would lead us to a point of epistemic fatalism, which in a world of action transforms into epistemic pragmatism.

But before considering the pragmatic turn, we should dispatch a potential misunderstanding. One might read the argument from epistemic attitude to entail that from *any* position of incomplete information we should be indeterminists. It would then be a sort of argument from ignorance, which argues that when we do not have *complete* information, indeterminism is the way to go. This would

trivialize the discussion since we can never actually occupy a position of *complete* information. Being an indeterminist would therefore *always* be more reasonable. This reading is a misunderstanding and does not follow from the argument presented. To see why, we should consider *when* the argument from epistemic attitude pushes us to be determinists from a position of incomplete information. With this aim, we will once again turn to our old-school determinists of the last century.

4.6 When being a determinist is reasonable: The old-school determinists again

Bell was right to proclaim Einstein's position reasonable, even though history has justified the critics. When quantum mechanics was first encountered at the turn of the last century, it was at a point in time where *deterministic* explanations reigned supreme. With the great explanatory success of classical mechanics, the more we learned about the world the more of it seemed explainable in deterministic terms. For centuries it had been the case that the more we learned about the world, the more deterministic it appeared. Although many complex phenomena were not yet subject to satisfying deterministic explanations, the *components,* whose combined interactions constituted these complex phenomena, seemed *individually* to be perfectly deterministic. In light of this, it was reasonable to assume that the more complex phenomena were fundamentally deterministic and that they would ultimately succumb to deterministic descriptions. It was reasonable to suspect that our failure to develop accurate deterministic descriptions of the complex phenomena was a consequence of our inability to accurately account for the chaotic interaction of the (deterministic) components: A consequence of our limited epistemic capacity. At that point in time, *determinists* were the ones who could satisfy the principle of a critical epistemic attitude.

Recall again the argument from epistemic attitude as it was put earlier:

1. As fallible creatures, we should apply a critical epistemological attitude when shaping and choosing our beliefs.
2. Having a critical epistemological attitude entails believing in a way so that we could learn (if not in practice, then in principle) *that* we are wrong *if* we are wrong.
3. In light of current best-explanation probabilistic models, only **indeterminists** can learn (in principle) that they are wrong if they are wrong.
4. In light of current best-explanation probabilistic models, we should be indeterminists.

This is how the argument goes today. But at the time of our old-school determinist, the current best-explanation models were *deterministic*, not indeterministic. At least up until the discovery of quantum mechanics. So, at the time of our old-school determinists, the third step looked different. Instead, it would have been:

3. In light of current best-explanation *deterministic* models, only **determinists** can learn (if not in practice then in principle) that they are wrong if they are wrong.

At the beginning of the last century, determinists could learn that they were wrong if they were wrong: Their future might produce best-explanations indeterministic models to replace the deterministic models. When the old-school determinists discussed why the indeterministic picture of quantum mechanics might be an incomplete or even paradoxical picture, as with the Einstein-Podolsky-Rosen hypothesis (Einstein et al. 1935), they were discussing and (implicitly) working towards an answer to the question of *how* they could learn that they were wrong *if* they were wrong. These critical discussions constituted a process that culminated in Bell's theorem and Bell-type experiments – and with those, an outcome that *did* show exactly what could be expected *if* our old-school determinists were in fact wrong. As such, our old-school determinists acted in accordance with the argument from epistemic attitude, as only the determinists at the time could learn (in principle) that they were wrong if they were wrong.

Consider again Einstein's comment to Born when he says that "No doubt the day will come when we will see whose instinctive attitude was the correct one" (Einstein et al. 1971, p. 149). 'No doubt' we will be able to see whose instinct was correct. This is the sound of someone who believes in a way so that they can *in principle* learn that they are wrong, if they are wrong. Because of the practical limitations of the time, the day was not yet ripe. It was still not possible *in practice*, but there was room for such a sight *in principle*.

Bell was right to call our old-school determinists reasonable, and his surprise at the outcome of Bell-type experiments is appropriate. But today, as we learn more about the world and as indeterministic models continue to be corroborated by experiments and outmatch deterministic models, even at the level of the components that compose the more complex higher-level phenomena, it is no longer the determinists that can satisfy the principles of a critical epistemological attitude. Only indeterminists can do so.

From what has been said so far, we seem to be in a position where we are hard-pressed to accept indeterminism either on epistemic principles with the argument from epistemic attitude or on empirical grounds with direct scientific realism. In either case, the pendulum swings towards indeterminism. But a third option remains. Perhaps the entire discussion seems unsatisfying for the reason that neither empirical insights nor principles of epistemic virtues should have any bearing on our conviction. This is a fair objection, which brings us to a third epistemological approach. It is this third approach that will constitute the final considerations of this paper.

4.7 Leaving metaphysics behind: Pragmatism proper and the pointless pragmatic

A common counterargument against any position on the matter at hand is that the question lies *outside the scope* of what we can justifiably say anything about. The truth of the matter is beyond the reach of human reason. The idea here is that the limits of reason, which make such questions unanswerable, make it reasonable to abandon the question altogether: The only reasonable position is one of pure and total agnosticism.

However, where the aim of our convictions cannot be truth, the hallmark for justification is *explanatory use*[35]. Although the *truth* of this metaphysical matter may be permanently out of reach, we must still interact with and understand the world. Even if our interaction with it is doomed to be rooted in some flawed, incomplete, understanding of it. This is the position of epistemic *pragmatism*, where the difference between *epistemic* and *ontological* indeterminism dissolves in a sea of irrelevance, and whose motto is ""We have to live today by what truth we can get today, and be ready tomorrow to call it falsehood" (James 1907, p. 150). For an epistemic pragmatist, what things look like at the edge of our epistemic horizon *is* how one takes them to be.

Yet here another objection may be raised. A particular species of pragmatist objects, that to be a pragmatist is to only ever be epistemically committed. As if pragmatists would never dare to speak of truth and have beliefs about *the world* as is. There is a subtle kind of hypocrisy at play in this branch of 'pragmatism'. This kind of pragmatist is certainly no determinist. He walks like an indeterminist, talks like an indeterminist, and interacts with the world like an indeterminist. *But he is not an indeterminist*. Or so he says. That is a step too far he claims. The position is a kind of epistemic limbo, where one dares

[35] Exactly what counts as 'use' as well as how general or individualized this notion should be understood is a topic for discussion beyond the scope of this paper.

not accept anything as real but think of everything as a fiction: A position Quine calls "a perverse one"[36] (Quine 1955, p. 241). This is the position of the pointless pragmatic: In practice to act in accordance with one belief about the world and yet refrain from believing this about the world. For if everything is *exactly as if* indeterminism is real, then what is the point of being a fence-sitting sceptic? I can think of only one reason: To preserve epistemic caution. But preserving caution does not require that we exclude having beliefs about the world. Caution is preserved when our beliefs are accompanied by the appropriate amount of uncertainty and a continued willingness to update our beliefs. In other words, so long that what we accept as true today is something we are ready to call false tomorrow.

It is difficult to imagine a greater handicap for any epistemology to suffer, than the inability to justify beliefs about *the world* in general. This is the perversion of epistemology that Quine warns about. Never ascribing full reality to things simply because there is some uncertainty in play: To have epistemic principles for justifications of beliefs without allowing these beliefs to be about the world and what is *real*. It is a disservice to the pragmatist tradition to impose this handicap upon it. Indeed, pragmatism proper does not imprison one's judgment in this epistemic limbo. It has clear principles for justification of beliefs, with which our beliefs can extend to be about the world and its ontology. If one accepts the pragmatist principles for justification and is asked if he believes the world is indeterministic, the current answer should be yes. The metaphysician may want to probe further and ask if the pragmatist believes if the world is merely *epistemically* indeterminate or *ontologically* indeterministic, but no pragmatist should answer yes to the first and no to the second, lest he be an epistemic pervert. As far as the pragmatist is concerned, the justification for one is the same as the other. He has already given his answer and the metaphysician is welcome to extrapolate that answer onto any impractical metaphysical subtleties he would like.

4.8 Concluding remarks

The arguments presented here constitute h a three-pronged line of reasoning for accepting ontological indeterminism. We are pressed to accept ontological indeterminism either on empirical grounds, on virtuous grounds, or from a position of epistemic pragmatism. I have argued that we should guard ourselves against the missteps of instinctive attitudes, and if we do so, find sufficient reasons for

[36] Quines himself writes that he "subscribe to determinism as fully as the quantum physicists will let me" (Quine 1981, p. 11). Expressing both the common default assumption of determinism, but also a willingness to let the scientific models have the last say, and update his belief in light of these models. Today, he would surely also let the neuroscientific models limit his assumption of determinism.

accepting ontological indeterminism. When we learn about the world we should apply a critical, but not dismissive, judgment – and be careful not to explain away what we *can* see by assuming that it *can't* be.

The advantage of the argument from epistemic attitude is that it provides us with a principle intrinsically tied to the universal virtue of self-criticism. And with this principle, we then have reasonable grounds for drawing conclusions in at least some cases of empirical underdetermination. Due to the holistic nature of beliefs, any time new information or new considerations problematize a given idea we hold as true, one of two options remain open to us: We can either change our mind about the belief in question, or we can change our mind about one or more of the surrounding assumptions. The big problem of empirical underdetermination is that there is no clear line for when we should adjust the auxiliary assumptions and when we should instead change our core beliefs. The principle of a critical attitude provides us with such a point of demarcation. It draws the line between when the continued adjustments of auxiliary assumptions remain justified (all other things equal), and when we should instead change our mind about the core belief in question. As it is derived directly from a clarification of what it is to be self-critical, the argument from epistemic attitude serves as a critical tool for choosing beliefs in the face of empirical underdetermination.

To answer the question in the title: We should be indeterminists when only the indeterminist position can satisfy the principle of a critical attitude. Similarly, the principle of a critical attitude also marks a clear point for when we should become determinists. If I am asked what I *feel* like the world is: Deterministic. If I am asked what I *think* the world is: Indeterministic. When I advocate indeterminism in the strong sense, it is not an attempt to justify my own intuitions. It is because my attempts to justify these intuitions have failed. I can find no critical justification for a belief in determinism. And know of no way to reach that conclusion without a too strong appeal to intuitions, habit of thought, or dogma.

5. Recursive Decisions, Indeterminism, and Luck

Abstract

Libertarian accounts of free will require that indeterminism is true – i.e. that at any instant more than one future is compatible with the laws of nature and the state of the universe at that instant. One of the primary challenges for such indeterministic accounts of free will is 'the problem of luck'. The problem arises from the idea that, if decisions are indeterminate, this entails an element of freedom-undermining luck in human agency. Here, I develop an account of free will where indeterministic agents can adjust their internal indeterminacy via *recursive decisions*: Decisions that reorder their motivational hierarchy. This kind of decision-making provides the agent with an aspect of control over their internal indeterminacy, which alleviates the problem of luck. Despite this, it is argued that the problem of luck ultimately remains unsolvable. If indeterminism is real we must accept innate randomness and luck as part of human agency.

5.1 Introduction

"Diana, a libertarian goddess in an indeterministic universe, wants to build rational, free human beings […]. Now, because Diana is a relatively typical libertarian, she believes that free decisions cannot be deterministically caused - even by something that centrally involves a considered judgment that it would be best to *A* straightaway […]. So Diana designs her agents in such a way that, even though they have just made such a judgment, and even though the judgment persists in the absence of biological damage, they may decide contrary to it.

Given Diana's brand of libertarianism, she believes that whenever agents freely perform an action of deciding to *A*, they could have *freely* performed some alternative action. She worries that her design does not accommodate this. Her worry, more specifically, is that if the difference between the actual world, in which one of her agents judges it best to *A* straightaway and then, at *t*, decides accordingly, and any possible world with the same past up to *t* and the same laws of nature in which he makes an alternative decision while the judgment persists is just a matter of luck, then he does not freely make that decision in that possible world *W."* (Mele 2006, p. 7-8)

This myth imagined by Mele illustrates one of the fundamental metaphysical challenges to libertarian notions of free will: The problem of luck[37]. The objection attacks libertarian free will – the notion that free will requires ontological indeterminism[38] - by claiming that indeterminism introduces a freedom-undermining element of luck (randomness) in the agent's decision-making and subsequent actions. The idea is that the randomness entailed by indeterminism undermines agentive control which is necessary for free will. As such, indeterminism threatens freedom of will. The problem comes in two versions – a soft version and a hard version. The soft version relates to indeterminism in the *execution of action* and the hard version relates to indeterminism in the *process of deciding*.

In this paper I will argue for two core claims: 1) The hard version of the problem of luck – the one illustrated by Mele's myth – is unsolvable. If indeterminism is true, we must accept that some internal randomness and luck is part of human agency. 2) Nevertheless, free agents have an important kind of control of their internal randomness, which alleviates the hard problem of luck. By directing decisions at their own motivational structure and future decisions they can influence the internal probabilities[39] of the agent. The capacity for this kind of decision-making constitutes an important kind of control and freedom in an indeterministic framework.

The primary aspects of the recursive account of freedom include that:

1. Freedom of the will is a *structural* matter – it depends on the structure and the *causal potential* of the agent, not the occurrence of any particular *event*.
2. The functional capacity that constitutes freedom of the will is the capacity for recursive decision-making – decisions directed at the agent's own motivational structure whereby the agent can effectively reorder their motivational hierarchy.
3. The capacity for recursive decision-making requires higher-order awareness of the agent's own motivational structure. Therefore, the agent's higher-order awareness of itself as a *motivated and agentive* creature is a necessary requirement for free will.

[37] Sometimes called the mind problem (Van Inwagen 2017).
[38] By indeterminism, I mean the idea that at any instant more than one future is compatible with the state of the universe at that instant and the laws of nature. Determinism is the negation of indeterminism, i.e. at any instant exactly one future is compatible with the state of the universe at that instant and the laws of nature (Mele 2006).
[39] Here and in general when speaking of 'probabilities' in this paper I am referring to the objective probabilities that are a product of the world's ontological indeterministic randomness. When considering probabilities from an epistemic perspective, however, I am open to the idea of 'imprecise probabilities' (Seamus 2019).

4. Freedom of the will does *not* depend upon how extreme or even the probabilities of competing first-order desires are – instead it depends upon the possible effect of recursive decisions on these probabilities: The potential *strength* of recursive decisions.
5. A free action does *not* require that a recursive decision about the effective motive has taken place prior to the action (see 1).
6. Freedom of the will is a graded phenomenon that can be meaningfully quantified as the product of the scope and strength of potential recursive decisions.

5.2 Indeterminism and luck

One of the primary objections against the idea of free will in indeterministic worlds is the problem of luck.

Formalized, the problem goes as follows:

- If an agent's acts are undetermined, then *how* they act on a given occasion is a matter of chance
- If how an agent act is a matter of chance then the agent does not have free will
- Ergo: if an agent's acts are undetermined, the agent does not have free will (van Inwagen 2017, see also Mele 2006)

There are two places where such indeterministic chance becomes problematic for freedom. The first is at the stage between decision and action – randomness in the *execution* of action. The second is at the stage between motivational structure and decision – randomness in the *choice* of action.

Randomness in the execution of action is what I call the 'soft' version of the problem of luck. Consider the example of a sniper who attempts to shoot his target but misses because some indeterministic events in his motor system lead to a twitch in his arm or a misevaluation of the trajectory of the bullet (Kane 1999, 2007). In such a case, luck clearly serves as an obstacle for the successful execution of an action, by preventing the agent from doing as they intend to. Although the role of indeterministic luck in the executive processes is an important problem for the analysis of *free action*, it is relatively unproblematic for questions of freedom of the will[40] (except on those accounts that hold that freedom in action and freedom of the will is identical[41]). Reducing randomness in the execution of action corresponds well with our everyday notion of skill: Someone who is skillful at a given

[40] See for instance Foot 1966 and Austin 1979.
[41] See Hobbes 1651.

action will have a high probability of executing it successfully. But lacking the skills to successfully execute certain actions is not the same as lacking free will. An unskilled agent may freely attempt to do what he lacks the skills to do successfully, but whether or not indeterministic events in his execution results in a successful or unsuccessful attempt, indeterministic interference at this point tells us nothing about whether the agent's will is free or not[42]. Similarly, a very skillful agent may occasionally fail under the best of circumstances due to innate randomness in the executive processes, but this also does not problematize our potential for free will. Even adults may stumble despite being skillful walkers. But the occasional stumble does not exclude that the person has free will. The successful execution of free *actions* may in part depend upon the internal randomness leading to one outcome instead of another, but this does not problematize free *will*.

Randomness in the choice of action, however, is where the problem of luck strikes at the heart of ontological indeterminism. This is the hard version of the problem which is illustrated by Mele's myth. Consider again our would-be assassin, but at a point in time before he has settled on the decision to kill his target. He has excellent reasons not to kill the target and has in fact recently come to the conclusion that it is best not to murder people at all. Nevertheless, because of some indeterministic events in his decision-making process, he decides (i.e. forms the intention) to assassinate his target. When he assassinates his target (or attempts to) he does it intentionally – but whether his will is free is precisely what the problem of luck brings into question. Even intentional actions may be unfree. This version of the problem, I contest, has no solution. It teaches us something about human agency that the world of philosophy has been reluctant to accept: That human agency and free will entails an aspect of unavoidable, innate, randomness. In an indeterministic world, the relationship between the motivational hierarchy of an agent and which motive actually becomes effective is a matter of chance: At any given moment, either of two (or more) competing motives have a certain probability of becoming effective (i.e. of moving the agent to action). If this was all there was to it, and these internal probabilities concerning which motive becomes effective was entirely out of the agent's hands, the notion of free will would indeed be in dire shape[43]. What a person needs is some aspect of control over these internal probabilities: A way to influence and shape them.

[42] Though it may problematize whether he intentionally did what he tried to do, in precisely the way that luck serves as a case of causal deviance. See for instance Mele & Moser 1994.

[43] An example of this can be found in Robert Kane's account of self-forming actions, where the indeterministic competition between motives constitutes the agent's 'efforts' and so whichever choice she makes she is considered free (Kane 1999). See also Chisholm 1966.

Before we get going we should note a common approach to the problem that will not be taken up here, which is the appeal to agent-causation (Clarke 2000, O'Connor 2000). I do not think such accounts provide a proper solution to the problem of luck for even if we grant that decisions are produced by agent-substance or emergent 'active power' this does nothing to undo the randomness of indeterministic relations or explain how these indeterministic relations come under the control of the agent[44]. Here, it is worth repeating Van Inwagen's recommendation to agent-causalists, when he says: "do not underestimate the power of the *mind* argument"[45] (Van Inwagen 2017, p. 165). The logic of agent-causation simply does not strike at the logic of the problem of luck (for a thorough analysis of this see Mele 2006, but for a reply see O'Connor 2007).

5.3 Recursive decisions: A tool against luck

Consider one of Diana's indeterministic agents, Sven. Sven's motivational structure consists of several different motives, many of them in conflict with each other. For instance, Sven has both the desire[46] to eat meat and the desire not to harm animals or contribute to their suffering. Sven has also judged it better to forgo his desire to eat meat in order not to contribute to the suffering of animals. One day, as Sven is walking through town, he gets hungry. Due to the indeterministic nature of his decision-making process, his desire to eat meat becomes effective and he forms the proximal intention[47] to eat meat. I.e. he decides against his better judgment to have a shawarma from the nearby shawarma place[48]. Two conflicting desires are occurrent, the desire to eat meat and the desire not to contribute to the suffering of animals. Both feed into the selection-process for action (the selection-process for which occurrent desire becomes *effective*), and each option has some probability between 0 and 1 of being selected (probabilities of precisely 0 and 1 are excluded, as this would constitute a case of determinism, and Diana's agents are fully indeterministic). Due to the indeterministic nature of this selection-process, the action motivated by the desire to eat meat is selected and thereby decided upon. When Sven is eating

[44] Although the O'Connor variety of agent-causation deserves special praise for his emergent picture of agency which avoids appealing to agent-substance (O'Connor 2000), his notion of emergent 'active power' does little to alleviate the problem of luck (See Mele 2006).
[45] Note that the mind-argument is what Van Inwagen calls the problem of luck.
[46] For the purpose of this paper, 'desire' is meant in the broad understanding of the term as it is commonly used in the literature, i.e. as pro-attitudes with a motivational component. (Davidson 1963, Frankfurt 1971).
[47] Intentions to act here-and-now, in contrast to distal intentions which are intentions to act at a later time (Mele 1992a).
[48] Scenarios as these – cases of weakness of will - are instantiations of luck-cases in worlds like Diana's. When all actions are luck-cases, all cases of weakness of will are cases of luck but not all cases of luck are cases of weakness of will. Whether the agent forms an intention in line with one's better judgment (rational) or an intention that conflicts with one's better judgment (irrational) is a matter of chance in Diana's world.

his shawarma, he is doing it intentionally and, supposedly, he thinks his decision was made freely. But Diana despairs, for she thinks that Sven's sense of freedom is illusionary, as a case of freedom-undermining luck has just occurred. I agree with Diana, for Sven remains ultimately a slave of desires which he has had no influence over. These desires and their corresponding weights are randomly distributed amongst him at the whims of nature, environment, or - in this case - Diana. Though they may be dynamic a changeable, for instance through repetition and the acquisition of new habits, Sven's desires and their weights are not of his own making.

What Sven lacks is the potential to shape his own motivational hierarchy; he lacks a mechanism for influencing his own internal probabilities. Suppose that Sven had the capacity to make decisions that target and alter his own motivational structure: The capacity for recursive decisions. With recursive decisions, an agent can then reorder his motivational hierarchy. When such a reordering is causally effective this amounts to directly influencing the probabilities for which motives become effective, although remains possible that the objective (ontological) probabilities do not perfectly mirror the intended reordering even after an effective recursive decision has taken place. It is if a person has such a mechanism to willfully and directly change their own internal probabilities, i.e. if the agent has the capacity for recursive decision-making, that a person has free will. Here it should be stressed that free will does not depend upon the probability distribution of first-order desires, regardless of how extreme or even they are. Free will depends instead on the *potential* for recursive decisions to affect these probability distributions of first-order desires.

Now, it is merely in *having* this capacity and not in *the use* of it that a person's will is free. So long as the capacity is there, the agent has *the potential* to shape his own motivational hierarchy. It is logically possible, however unlikely, that there should exist an agent with the capacity for recursive decision-making without ever making use of it. For example, a person might upon reflection find that the hierarchy of their desires is exactly as they want it to be, and so they choose not to change anything about it. Such a person has free will even if they never make a recursive decision in their life. This does not amount to saying that none of his actions are free if no recursive decision takes place. On the contrary, all or almost all of the agent's actions will be quite free even if no recursive decision occurs. Freedom of the will is ultimately a *structural* matter: It concerns the agent's causal *potential*. Specifically, the causal potential of recursive decisions, i.e. the extent to which such recursive decisions can change the objective probabilities that constitute the motivational structure of the agent. A person may neglect the use of their freedoms, but this is not the same as being without them.

If Sven has recursively reordered his motivational structure, it is no longer the case that he is at the whims of desires whose weights are not of his own making. They are - at least in part. In a certain sense, he has created his own luck, regardless of which decision he ends up making. He is to a certain degree – but to a degree only – the craftsman of his own inner roulette wheel. By influencing his internal probabilities via recursive decision-making, Sven himself becomes an important part of the multifaceted causes of his motivational structure[49]. Reordering one's own hierarchy clearly requires that the agent perceives themselves as an agent with such a motivational hierarchy and the awareness both that these motives can be conflicting and that they can change. Recursive decisions thus depend upon this kind of higher-order awareness[50]. It is only with this awareness and the effective reordering of their own motivational structure that an agent can make decisions about *what kind of agent* he becomes. Here, we also see why the mere potential for changing inner probabilities via repetition and the acquisition of new habits is itself insufficient for free will. Since creatures with this kind of higher-order awareness clearly can make recursive decisions (at least phenomenologically), this class of decisions would be epiphenomenal if any actual change in objective probabilities was entirely dependent on some further repetitions beyond the decision itself.

Such recursive decisions can be said to be done for a reason in the Melean sense, where this does not necessarily require any belief-component or reflective deliberation, and where actions done for bad reasons (even from the agent's point of view) as well as intrinsically motivated actions that are done for no *further* reason counts as 'being done for a reason' (Mele 1992b). A recursive decision can thus take place as a result of the agent being confronted with new reasons, gaining new desires[51], or forming new beliefs – perhaps as a result of self-evaluation. Sven might, for instance, upon learning about the cruelties imposed on animals in the food sector form the second-order desire not to be moved by his taste for meat, which triggers a decision to reorder his motivational hierarchy and thereby reduce the probability that his desire to eat meat will become effective in the future.

There are at least two ways in which a recursive decision might modify the probability for which desire becomes effective. The first is by a reduction of the intensity of desire and the second is by an

[49] The capacity for recursive decision-making can also be mapped onto a compatibilist framework. Here, however, the focus is to consider this account within an incompatibilist framework and explore how it relates to the problem of luck.

[50] Though this does not require that the agent has an accurate awareness of their motivational structure, only that they perceive themselves as the kind of thing that has motives and a motivational structure.

[51] This is not to say that recursive decisions themselves *create* new desires. Though the creation and destruction of new desires is an interesting part of the human condition, how this comes about is not something to be explored in detail here, as it is inconsequential for our current analysis.

increase in the amount of willpower. First of all, recursively reordering the motivational hierarchy might simply affect the desires themselves. If the agent, for instance, makes a decision that recursively reduces the probability that a certain desire becomes effective (such as pre-emptively deciding against acting upon it when the opportunity presents itself), this might simply result in that desire being felt comparatively less intensely at the moment of action. Neurally, this could be realized by neuromodulatory effects on the representations of the desire, such that the representation activates less easily from upstream-stream influences (i.e. relevant stimulus) and such that, when it activates, it results in smaller down-stream effects (i.e. on the selection-process for action whose inputs are the competing desires and where either the threshold for acting is met or not)[52]. Essentially reducing the desire, which is to say, reducing how much desire the agent feels for a given 'option' when confronted with the opportunity to choose it. The second way a recursive decision might influence the internal probabilities is by modifying how much 'synchronous self-control', i.e. 'willpower' (Kennett & Smith 1996, Mele 1997, Holton 1999, Sripada 2014), the agent is likely to exert when the desire is activated. For instance, by a re-sculpting of the connections between neural representations of desires and the circuitry relevant for synchronous self-control (for instance particular pre-frontal regions (Heatherton & Wagner 2011)), such that the latter is more sensitive to the activation of the former – thus increasing the chance that effortful synchronous self-control will be attempted and increase the degree of such efforts. Such dynamic re-sculpting of the relevant neural circuits would be cases of 'criterial causation' and could occur via dynamic synaptic resetting as outlined by Peter Tse (Tse 2011). Although this kind of decisive and agentive reordering of the motivational hierarchy would be a contrast to the kind of motivational restructuring that takes place entirely through repetition, such habitual learning through repetition may very well accompany and either supplement or work against the effects of a recursive decision.

Nevertheless, if a person has recursively reordered the hierarchy of his motives so that one motive outweighs the other and thereby has a higher probability of becoming effective, it must remain possible that the lower-ranking motive will become effective simply due to internal randomness in the decision-making process. This must be the case to keep in line with Diana's libertarian requirement that free decisions cannot be deterministically caused. However, there is no theoretical limit for how small such a probability can be (only that cannot be *infinitely* small or 0). For instance, it is possible that Sven

[52] Neuromodulatory effects of agency have for instance been demonstrated on visual and attentional processes (Loyola-Navarro et al. 2020). Interestingly, the authors suggest modulatory effects on dopamine, acetylcholine, and noradrenergic circuits – circuits which have also been shown to be of key importance for the representation of desires (or of 'wanting') (Daw & Shohamy 2008, Kim 2013).

recursively modifies his motivational structure such that the probability that his desire to eat meat is so astronomically small that even if Sven has to decide between meat and vegetables several times a day for his entire life, there is less than a 0.1% chance that the desire to eat meat will become effective even once. It is possible, that the probability that certain desires become effective is so small that it becomes what ordinary non-metaphysicians would call *practically* determined[53]. But recursive decisions cannot make a fundamentally indeterministic universe deterministic. Even with recursive decision-making, we are then, to a certain degree, always at the mercy of chance. But this degree is within the agent's grasp to change. And recursive decisions are the agent's tool for influencing these inner probabilities: A tool against the randomness within us.

5.4 Luck again: Biting bullets and further consequences

Would the above account of recursive decision-making satisfy Diana's worries? I suspect not. Perhaps neither Diana nor Mele will be satisfied that recursive decision-making solves the hard problem of luck. Even if they agree that that recursive decision-making does grant the agent an important kind of control and influence over their own internal randomness, they might still object to it as a genuine solution for the reason that randomness and luck is still part of the picture in at least two ways. First of all, even if the motivational hierarchy of the agent has been recursively modified, it is still a matter of chance whether the agent acts one way or another at the moment of action. After all, recursive decisions only influence the *probabilities* that certain desires will become effective. Even if Diana imbues her agents with the capacity for recursive decision-making, she is left with an agent who might act contrary to his best judgment at time *t* even with the same past up to *t* and the same laws of nature. And so, Diana's dilemma remains. But even if Diana accepts that the problem of luck is alleviated by the agent having some degree of control over his internal probabilities, this may do little to alleviate her worries, for she notices that the problem reappears at a new level. Whether the agent decides to recursively reorder his motivational hierarchy, as well as which order he decides to reorder it to, seems now a matter of chance. Recursive decisions are also part of the indeterministic world and as such subject to randomness.

This regress is unavoidable. What Diana sees clearly is of course, as Hume put it, that "tis impossible to admit of any medium betwixt chance and absolute necessity" (Hume 1739, p. 124). If ontological indeterminism is true, "…the impossibility of ultimate self-determination as to desire is a

[53] What Doyle calls 'adequate determinism' (Doyle 2013)

simple conceptual truth…" (Strawson 1986, p. 42). Simply put, the hardest version of the problem of luck cannot be undone within an ontologically indeterministic framework. It is an unavoidable part of it and has no strong solution. If we insist that free will excludes deterministically caused choices, as Diana thinks, the problem of luck confronts us with a conceptual choice. We must either accept that the concept of free will is an innate contradiction, a *contradictio in terminis,* or - if we want to continue talking intelligibly of free will without turning to compatibilism - we must reconceptualize it in a way that accepts this randomness as part of it. Taking this choice seriously is by no means a trivial matter in philosophy, and accepting this kind of reconceptualization may seem to bite a bullet too big to swallow. Van Inwagen for instance arrives at this junction without taking either position. He thinks that there must be something wrong with the problem of luck but he hasn't "the faintest idea what the nature of the error is" (Van Inwagen 2017, p. 165). A conclusion to a thorough reflection on the problem that is as remarkably honest as it is problematic: To have an argument with which one can find no error after thorough investigation - and yet to reject the conclusion.

Instead of despairing at the problem of luck or leaving it at the doorstep of mystery, we should take its lessons to heart. What the problem shows is that if ontological indeterminism is true, we must accept internal randomness as part of human agency. This embrace of luck can also be found in Robert Kane's account of self-forming actions, where he dissolves the line between indeterminism and agent when he says that there is no point at which the agent stops and "chance 'takes over'" (Kane 1999, p. 232).

Although the fundamental randomness that is part of us cannot be eliminated, it can be increased or decreased by influencing the probabilities of our internal dice-game if the agent has the capacity for recursive decisions. One consequence of this view is that freedom of the will can sensibly be quantified as a *matter of degrees*. The greater the influence our recursive decisions can have on our internal probabilities, the freer our will is. As already noted free will does not depend upon the probability distributions of the competing desires but on the potential effect of recursive decisions upon these probabilities. In other words, it is not the probability distributions of the different desires in the motivational hierarchy that constitute the degree of freedom but rather it is the *potential effectiveness* of recursive decisions to change these internal probabilities that constitutes the degree of free will the agent has. How much of an effect a recursive decision *potentially* has, in turn, depends on the structure of the agent – just as the potential speed of a projectile depends on the structure of the device that launches it (famously, the Trebuchet can launch a similarly weighted projectile much further than the Onager Catapult). An agent whose recursive decisions can have a great influence on his motivational

structure and the weights of his desires is more free than someone whose recursive decisions only have a weak influence. Both agents are free, but one more so than the other. An increase in freedom may come through learning via introspection, e.g. to the extent that self-efficacy beliefs about free will and recursive capabilities can increase the effects of recursive decisions (Bandura 1977, 2006), or it may grow through training whereby an agent that actively engages in recursive decision-making can strengthen the influence of any future recursive decisions. In this way, recursive decision-making is also in a sense a skill that one can improve. Free will is then something that can be *learned and trained* over the course of an agent's life. In any case, the stronger the potential effects of recursive decisions, the freer the agent: The better equipped he is to take control of the randomness within him. Free will is then more appropriately conceptualized as a graded phenomenon. If we similarly hold that the notion of *agency* corresponds to the extent that an agent can *determine* whether a given action occurs or not and since *complete* self-determination is impossible within an indeterministic ontology, then agency too is something graded. The quantification of freedom of the will can be seen as a product of the *strength* and *scope* of the agent's recursive decision-making. Where strength indicates the size of the effect the agent's recursive decisions can have on his internal probabilities, i.e. by how much he can increase or decrease his internal probabilities by recursively reordering his motivational hierarchy, and scope indicates the proportion of the agent's total desires that he can recursively influence. The maximally free agent is one who can recursively reorder every one of their desires and for whom their recursive decisions have a maximally strong effect (whatever this limit may be).

A second consequence that follows the graduation of free will is that free will can vary across time and context within the same person. It may be, for instance, that the influence of recursive decisions does not extend to certain desires and that these desires have a very high probability of becoming effective when confronted with their object. In other words, it is possible that recursive decisions can only affect a subset of the agent's total motives, and that some motives outside of this subset are very likely to become effective. Consider, for example, the case of drug addiction. If the drug addict is one of Diana's agents, there is of course still some (vanishingly small) chance that he will not take his drug of choice when presented with the opportunity. But furthermore, it may be that the drug addict cannot recursively affect his desire to take his drug of choice – his higher-order desire remains, but his capacity for recursively reordering his desire to take the drug is reduced to nil. The probability for whether he takes his drug or not in situations where he is confronted with it is then entirely out of the addict's hands. By handicapping his capacity to recursively influence his desire to take drugs the addiction robs the agent of his free will in situations where this desire is confronted with its object.

However, it is possible that the agent can nevertheless recursively reorder the hierarchy of other motives within him. It is then not in every situation that the addict lacks free will. Only when he acts on a desire that is outside the scope of his recursive decision-making is he acting without free will. It is only within particular contexts that the addict's freedom of will suffers. However, it is also possible that a desire which is at one time outside the scope of the agent's recursive decisions may at a later point fall within this scope and be subject to recursive influence. Similarly, it may be that the strength of an agent's recursive decisions – especially as regards certain desires – may increase or decrease over time. For instance, when children first acquire the capacity for recursive decision-making, such decisions may only have weak effects on their motivational hierarchy. But with time the strength of a person's recursive decisions may grow (or diminish). Both the scope and the strength of an individual's recursive decision-making may change and develop over time.

What we are left with is then a changing, developing, and growing kind of free will. One that changes over the course of an agent's lifetime and can be quantified in the manner outlined above. Agents with the capacity for recursive decisions have certainly gained an extra degree of freedom that is not entailed by mere indeterministic decision-making since they have gained an aspect of control over their own internal probabilities. Whether this is enough to entirely alleviate Diana's worries is for her to decide, but in any case, it is as much freedom of the will as one can hope for in an indeterministic reality.

6. Concluding Remarks and Future Prospects

At the beginning of this thesis, I asked how probabilistic uncertainties in the mechanisms underlying intentional action should affect our understanding of free will. The objective of the thesis was to provide an account of free will that was sensitive to the folk-conception of free will, the probabilities reflected in neuroscientific models, and the problem of luck. In doing so, the thesis aimed to contribute to the folk-psychology of free will, general epistemology but especially as it pertains to the interpretation of probabilities, and finally to the general free will debate by providing an account for agentive control in light of luck whereby the agent can influence their internal probabilities via recursive decision-making. In the four papers that make up this dissertation, four particular questions related to the research question was addressed:

1. What is the folk-concept of free will?
2. Does belief in physicalism drive different free will conceptions?
3. Are there good reasons to interpret the probabilities of scientific models in terms of ontological indeterminism?
4. What kind of luck-reducing control might indeterministic agents have over their own internal probabilities?

The first two questions were addressed by running a set of experiments. The first experiment aimed to assess the folk-concept of free will. Two novel additions were included in the experimental design. First of all, by including vignettes that framed indeterminism as the presence of probabilistic chance, the design accounted for the possibility that laypeople have hard incompatibilist intuitions whereas previous designs have assumed that the ordinary concept of free will must be either compatibilist or incompatibilist. Secondly, the design included controls for understanding errors such as bypassing and intrusion effects, to ensure that participants understood the philosophically relevant concepts as intended, i.e. in the philosophically relevant way. These novelties were carried over in the follow-up experiment that looked into whether beliefs in physicalism contra substance dualism lead to different free will conceptions. Our findings support the idea that folk-concepts are incompatibilist, but not hard incompatibilist. And though it appeared that more physicalist beliefs drove more incompatibilist free will conceptions, this effect was shown to in fact be mediated by physicalists being more likely to interpret deterministic descriptions as entailing epiphenomenalism. Nevertheless, when determinism was

interpreted as philosophers intend it, we found that whether people were physicalist or substance dualist their free will conception was incompatibilist.

These two experiments have at least two significant limitations, however. The first and perhaps most pressing limitation that invites further research, is that the experiments here only employed non-moral vignettes. It has been shown that the emotional salience of the vignettes affects laypeople's free will conceptions and that more emotionally salient vignettes drive more compatibilist conceptions. A promising venture for future research is to investigate if this effect of emotional salience on the folk-conception of free will remains when controlling for the understanding errors as we have done here. The findings from such experiments, that apply moral vignettes with the appropriate controls, might further solidify – or problematize – the findings of this thesis. A second and significant limitation is that the populations tested in the two experiments here were western-based populations. It is an open question worth investigating whether or not the results will be cross-culturally validated. Since some cross-cultural differences have already been demonstrated in the literature, it will be of interest for further investigations whether free will conceptions genuinely vary across cultures or if any such difference is driven by variance in the common understanding errors.

The third question was addressed in the third paper. It was argued that regardless of which epistemic approach we adopt, direct scientific realism, virtue epistemology, or epistemic pragmatism we have sufficient reason to accept ontological indeterminism in light of best-explanation probabilistic models. Particular emphasis was put on the virtue epistemological approach and the virtue of self-criticism. It was argued that in order to be genuinely self-critical, we must be able to answer what could, in principle if not in practice, change our view. In light of best-explanation probabilistic models, it was argued that only indeterminists could satisfy the principle of a critical attitude. Thus, we should be ontological indeterminists.

The fourth question was addressed in the fourth and final paper of this thesis. Although it was conceded that the hard problem of luck had no perfect solution, it was argued that this confronted us with a conceptual choice: We must either give up the idea of freedom in ontologically indeterministic worlds or we must accept an aspect of randomness as part of free will. It was argued that, although an aspect of luck remains, indeterministic agents with the capacity for recursive decision-making gain an important kind of control over their own internal probabilities – a kind of control that constitutes freedom of the will. Within this framework, freedom of the will was conceived as a graded phenomenon tied to the causal potential of an agent's recursive decisions. Although it is argued that the capacity for recursive decision-making constitutes freedom of the will in indeterministic agents, the question

remains whether human beings actually *have* this capacity or not. Translating the recursive framework into a concrete empirical model with specific hypotheses to be put to the test would be an exciting future prospect for those who agree that this capacity would constitute free will.

One final future prospect that deserves mention is that the discussion so far has omitted any comments on *moral responsibility*. Indeed, moral responsibility is at the heart of questions and discussions about free will. It will no doubt be a matter of discussion, whether the account of free will as the capacity for recursive decision-making salvages any justification for moral responsibility, as well as what parts – and which parts, if any, of our ordinary moral intuitions should be revised in light of this account.

Supplementary Notes

Supplementary notes 1:

For the high-stake version of the deterministic vignette, the last paragraph was changed to the following:

"For example, one day John decides to marry his fiancée Laura. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it *had to happen* that John would decide to marry Laura at that point in time. If time was turned back over and over again to the exact same moment just before John made his decision, then it had to happen that he would decide to marry Laura at that moment every time."

For the high-stake version of the indeterministic vignette, the last paragraph was changed to the following:

"For example, one day Finn decides to marry his fiancée Patricia. Like everything else, this decision was not completely caused by what happened before it. So, if everything in this universe was exactly the same up until Finn made his decision, then it *did not have to happen* that Finn would decide to marry Patricia at that point in time. If time was turned back over and over again, to the exact same moment just before Finn made his decision, then it did not have to happen that he would decide to marry Patricia at that moment every time."

For the high-stake version of the probabilistic, the last paragraph was changed to the following:

"For example, one day Martin decides to marry his fiancée Julia. Like everything else, this decision was matter of probabilities. So, if everything in this universe was exactly the same up until Martin made his decision, then *there was a certain probability* that Martin would decide to marry Julia at that point in time. If time was turned back over and over again, to the exact same moment just before Martin made his decision, then some of the times he would decide to marry Julia at that moment, and some of the times he would decide not to marry Julia at that moment.

Supplementary notes 2:

The six understanding questions:

1. **Decisions:** "In Universe [x], do a person's decisions have an effect on what they do?"
2. **Desires:** "In Universe [x], do a person's desires have an effect on what they do?"
3. **Beliefs:** "In Universe [x], do a person's beliefs have an effect on what they do?"
4. **Control over:** "In Universe [x], does a person have control over what they do?"
5. **Past:** "In Universe [x], does the past have an influence on what a person does?"
6. **Could have done otherwise:** "In Universe [x], could a person have decided to do otherwise?"

The scale for answering questions 1-5:

- Yes
- No
- Don't Know

The scale for answering question 6:

- Yes, Even if she had the exact same wants and beliefs
- Only if her wants or beliefs had been different
- No, not even if her wants and beliefs had been different
- Don't Know

For analysis, the 'past' and 'control over' questions were excluded because they have no clear right or wrong answer for the indeterminism and probabilistic vignettes.

Supplementary notes 3:

The understanding questions were scored in the following way:

For the 'decisions', 'desires', and 'beliefs' questions, the correct answer to these questions in all universes was yes, indicating the presence of mental causation. A correct answer to either of these questions was coded as +1 with an incorrect answer or an answer of 'don't know' being coded as -1.

The 'could have done otherwise' question was the most important of the control questions, as getting this wrong indicates a different understanding of the metaphysics described in the vignettes than what philosophers have in mind. The correct answer to this question was "Only if her beliefs and desires had been different" for the determinism vignette, and "yes, even if her beliefs and desires were exactly the same" for the indeterminism and probabilistic vignettes. The "no, not even if her beliefs and desires had been different" reply was wrong in all cases, as it indicates an epiphenomenal reading of the vignettes (the bypassing effect). The 'yes, even if her beliefs and desires were exactly the same' was wrong in the deterministic vignettes, indicating an indeterminist reading (the intrusion effect). An answer of 'don't know' was also coded as wrong, as it did not indicate that the participant clearly understood the vignette in the intended way.

The 'could have done otherwise' was then coded as either +3 (right answer) or -3 (wrong answer). This way the 'could have done otherwise' question served as a threshold, as no response that got the 'could have done otherwise' question wrong could have a score above 0, and a response to the control questions that scored 0 or lower were indicative of the presence of either the bypassing effect or the intrusion effect.

Supplementary notes 4:

The two understanding questions and their possible answers are as follows:

1. "Click the element or elements that have an effect on what a person does, in the world you just read about."
    - Decisions
    - Thoughts
    - Eye color
    - Feelings
    - None of the above
    - Don't Know
2. "In the world you just read about, if we go back in time to *just* before John made the decision to marry his girlfriend, could he have made a different decision?"
    - Yes
    - No
    - Don't Know

Supplementary notes 5:

The free will rating question and the possible answers:

1. "In the world you just read about, is it possible for a person to have free will?"
    - Yes, very certain
    - Yes, somewhat certain
    - Yes, very uncertain
    - Don't Know
    - No, very uncertain
    - No, somewhat certain
    - No, very certain

Supplementary notes 6:

The five dualism questions:

1. The fact that we have souls that are distinct from our material bodies is what makes humans unique.
2. Each person has a non-physical essence that makes that person unique.
3. The human mind cannot simply be reduced to the brain.
4. The human mind is more than just a complicated biological machine.
5. Human action can only be understood in terms of our souls and minds and not just in terms of our brains.

The scale for answering the dualism questions:

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly disagree

Bibliography

Abbott L. F., Varela J. A., Sen K. & Nelson S. B. (1997). Synaptic Depression and Cortical Gain Control *Science*, 275(5297):220-224.

Adler J. (2004). Reconciling Open-Mindedness and Belief. *Theory and Research in Education*, 2(2):127-142.

Agranov M. & Ortoleva P. (2017). Stochastic Choice and Preference for Randomization. *Journal of Political Economy*, 125(1):40-68.

Alexander J. & Weinberg J. (2007). Analytic Epistemology and Experimental Philosophy. *Philosophy Compass*, 2(1):56-80.

Almeida M. & Bernstein M. (2003). Lucky Libertarianism. *Philosophical Studies*, 113(2):93-119.

Alvarez O., Gonzalez C. & Latorre R. (2002). Counting Channels: A Tutorial Guide on Ion Channel Fluctuation Analysis. *Advances in Physiology Education*, 26(1-4):327-341.

Andrews F. M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly*, 48(2):409-442.

Arndt M., Nairz O., Voss-Andrea J., Keller C., Van der Zouw G. & Zellinger A. (1999). Wave-Particle Duality of C60 Molecules. *Nature*, 401(6754):680-682.

Auger C. & Marty A. (2000). Quantal Currents at Single-site Central Synapses. *Journal of Physiology*, 526(Pt 1):3-11.

Austin J. L. (1979). ”Ifs and Cans” in J. L. Austin, J. O Urmson & G. J Warnock (eds.), *Philosophical Papers*. Clarendon Press.

Bandura A. (1977). Self-Efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review*, 84(2):191-215.

Bandura A. (2006). Toward a Psychology of Human Agency. *Perspectives on Psychological Science*, 1(2):164-180.

Baumeister R. (2008). Free Will in Scientific Psychology. *Perspectives on Psychological Science*, 3(1):14-19.

Bell J. S. (1964). On the Einstein Podolsky Rosen Paradox. *Physics*, 1(3):195-200.

Bering J. M. (2006). The Folk Psychology of Souls. *Behavioral and brain Sciences*, 29(5):453-462.

Bering J. M., & Bjorklund D. F. (2004). The Natural Emergence of Reasoning About the Afterlife as a Developmental Regularity. *Developmental Psychology*, 40:217-233.

Berinstein M. (1995.) Kanean Libertarianism. *Southwest Philosophy Review*, 11(1):151-157.

Berkeley G. (1710). *A Treatise Concerning the Principles of Human Knowledge*. Dover Philosophical Classics.

Bernstein J. (1990). *Quantum Profiles*. Princeton University Press.

Björnsson G. (2014). “Incompatibilism and ‘Bypassed’ Agency” in A. Mele (ed.), *Surrounding Free Will*. Oxford University Press.

Björnsson G. & Pereboom D. (2016). “Traditional and Experimental Approaches to Free Will and Moral Responsibility” in J. Sytsma & W. Buckwalter (eds.), *Companion to Experimental Philosophy*. Blackwell.

Bloom P. (2004). *Descartes' Baby*. Basic Books.

Bohm D. (1957). *Causality and Chance in Modern Physics*. Van Nostrand Reinhold Company.

Bok H. (1998). *Freedom and Responsibility*. Princeton University Press.

Born M. (1949). *Natural philosophy of cause and chance*. Clarendon Press.

Bourget D. & Chalmers D. (2013). What do Philosophers Believe? *Philosophical Studies*, 170(3):465-500.

Brandon R. N. & Carson S. (1996). The Indeterministic Character of Evolutionary Theory: No "No Hidden Variables Proof" but No Room for Determinism Either. *Philosophy of science*, 63(3):315-337.

Brembs B. (2011). Towards a Scientific Concept of Free Will as a Biological Trait: Spontaneous Actions and Decision-Making in Invertebrates. *Proceedings of the Royal Society B,* 278(1707):930-939

Britten K. H., Newsome W. T., Shadlen M. N., Celebrini S. & Movshon J. A. (1996). A Relationship Between Behavioral Choice and The Visual Responses of Neurons in Macaque MT. *Visual Neuroscience,* 13(1):87-100.

Broad C. D. (1952). "Determinism, Indeterminism and Libertarianism" in C. D. Broad *Ethics and the History of Philosophy.* Routledge & Kegan Paul.

Butterfield J. (2012). Laws, Causation and Dynamics at Different Levels. *Interface focus,* 2(1):101-114.

Carelli P. V., Reyes M. B., Sartorelli J. C. & Pinto R. D. (2005). Whole Cell Stochastic Model Reproduces the Irregularities Found in the Membrane Potential of Bursting Neurons. *Journal of Neuroscience,* 94(2):1169-1179.

Carnap R. (1959). "Psychology in Physical Language" in A.J. Ayer (ed.), *Logical Positivism*. The Free Press.

Caruso G. (2012). *Free Will and Consciousness*. Lexington books.

Caruso G. (2016). Free Will Skepticism and Criminal Behavior: A Public Health-Quarantine Model. *Southwest Philosophy Review,* 32(1):25-48.

Caruso G. (2019a). "Free Will Skepticism and Its implications: An Argument For Optimism" in E. Shaw, D. Pereboom & G. D. Caruso (eds.), *Free Will Skepticism in Law and Society*. Cambridge University Press.

Caruso G. (2019b). Skepticism About Moral Responsibility. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2018/entries/skepticism-moral-responsibility/

Cashmore A. (2010). The Lucretian Swerve: The Biological Basis of Human Behavior and the Criminal Justice System. *PNAS,* 107(10):4499-4505.

Cassirer E. (1956). *Determinism and Indeterminism in Modern Physics*. Yale University Press.

Chisholm R. (1964). "Human Freedom and the Self". Reprinted in *Agency and Responsibility* (2001) By L. W. Ekstrom (ed.). Routledge.

Chisholm R. (1966). "Freedom and Action" in K. Lehrer (ed.), *Freedom and Determinism*. Random House.

Churchland P. M. (1984). *Matter and Consciousness*. MIT Press.

Churchland P. S. (1986). *Neurophilosophy*. MIT Press.

Clarke R. (1996). Agent Causation and Event Causation in the Production of Free Action. *Philosophical Topics,* 24(2):19-48.

Clarke R. (2000). Modest Libertarianism. *Philosophical Perspective,* 14(1):21-45.

Clarke R. (2003). *Libertarian Accounts of Free Will*. Oxford University Press.

Code L. (1987). *Epistemic Responsibility*. University Press of New England and Brown University Press.

Colquhoun D. & Hawkes A. G. (1995). "The Principles of the Stochastic Interpretation of Ion-Channel Mechanisms" in Sakmann B, Neher E. (eds.), *Single-Channel Recording*. Springer.

Cova F., Bertoux M., Bourgeois-Gironde S. & Dubois B. (2012). Judgments about Responsibility and Determinism in Patients with Behavioural Variant of Frontotemporal Dementia: Still Compatibilist. *Consciousness and Cognition,* 21(2):851-864.

Dangerfield C. E., Kay D. & Burrage K. (2010). Stochastic Models and Simulation of Ion Channel Dynamics. *Procedia Computer Science,* 1(1):1587-1596.

Davidson D. (1963). Actions, Reasons and Causes. *The Journal of Philosophy,* 60(23):685-700.

Davidson D. (1971). "Agency" reprinted in D. Davidson's *Essays on Actions and Events*. Clarendon Press.

Davidson D. (1973). "Freedom to Act" reprinted in D. Davidson's *Essays on Actions and Events* (1980). Clarendon Press.

Daw N. D. & Shohamy D. (2008). The Cognitive Neuroscience of Motivation and Learning. *Social Cognition*, 26(5):593-620.

Dayan P. & Abbott L. F. (2001). *Theoretical Neuroscience*. MIT Press.

De Brigard F., Mandelbaum E. & Ripley D. (2008). Responsibility and the Brain Sciences. *Ethical Theory and Moral Practice*, 12(5):511-524.
De Finetti B. (1974). *Theory of Probability*. Wiley.
Dean A. F. (1981). The Variability of Discharge of Simple Cells in the Cat Striate Cortex. *Experimental Brain Research*, 44(4):437-440.
Del Cul A., Baillet S. & Dehaene S. (2007). Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness. *PLOS Biology*, 5(10):e260.
Dennett C. D. (2013). A Dialogue on Free Will. *Methodé*, 2(3):46-48.
Descartes R. (1641). *Meditations on First Philosophy*. Reprinted in J. Cottingham, R. Stoothoff & D. Murdoch (eds.), The Philosophical Writings of Descartes. Cambridge University Press.
Diba K., Lester H. A. & Koch C. (2004). Intrinsic Noise in Cultured Hippocampal Neurons: Experiment and Modeling. *Journal of Neuroscience*, 24(43):9723-9733.
Dorval A. D. Jr. & White J. A. (2005). Channel Noise is Essential for Perithreshold Oscillations in Entorhinal Stellate Neurons. *Journal of Neuroscience*, 25(43):10025-10028.
Doyle R. O. (2013). "The Two-Stage Model to the Problem of Free Will" in Suarez A. & Adams p. (eds.), *Is Science Compatible with Free Will?*. Springer.
Dudman J. T. & Nolan M. F. (2009). Stochastically Gating Ion Channels Enable Patterned Spike Firing Through Activity-Dependent Modulation of Spike Probability. *PLoS Computational Biology*, 5(2):e1000290.
Duhem P. (1954). *The Aim and Structure of Physical Theory*. Princeton University Press.
Einstein A. (1936). Physik und Realität. *Journal of the franklin institute*, 221(3):313-347.
Einstein A., Born M. & Born H. (1971). The Born-Einstein Letters 1916-1955, first published in *The Born-Einstein Letters* (1971). Macmillian press.
Einstein A., Podolsky B. & Rosen N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical review*, 47(10):777-780.
Ekstrom L. W. (2002). "Libertarianism and Frankfurt-style Cases" in Kane R. (ed.), *The Oxford Handbook of Free Will*. Oxford University Press.
Ermentrout G. B., Galán R. F. & Urban N. N. (2008). Reliability, synchrony and noise. *Trends in Neuroscience*, 31(8):428–434.
Faisal A. Selen, L. & Wolpert D. (2008). Noise in the Nervous System. *Nature Review Neuroscience*, 9(4):292–303.
Feltz A. & Cokely E. (2019). Extraversion and Compatibilist Intuitions: A Ten-year Retrospective and Meta-analyses. *Philosophical Psychology*, 32(3):388-403.
Feltz A., Cokely E. & Nadelhoffer T. (2009). Natural Compatibilism versus Natural Incompatibilism: Back to the Drawing Board. *Mind and Language*, 24(1):1-23.
Feltz A. & Cova F. (2014). Moral Responsibility and Free Will: A Meta-Analysis. *Consciousness and Cognition*, 30(1):234-246.
Fine A. (1986). *The Shaky Game: Einstein, Realism and The Quantum Theory*. University of Chicago Press.
Fischer M. J. (1994). *The Metaphysics of Free Will*. Wiley-Blackwell.
Fischer M. J. & Ravizza S. J. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.
Foot P. (1966). "Free Will as Involving Determinism" in B. Berofsky (ed.), *Free Will and Determinism*. Harper & Row.
Forstmann M. & Burgmer P. (2015). Adults are Intuitive Mind-Body Dualists. *Journal of Experimental Psychology: General*, 144(1):222-235.
Forstmann M. & Burgmer P. (2018). A Free Will Needs a Free Mind: Belief in Substance Dualism and Reductive Physicalism Differentially Predict Belief in Free Will and Determinism. C*onsciousness and Cognition*, 63:280-293.
Frank H. T. & Toshinori H. (2015). "Models of Ion Channel Gating" in Zheng J. and Trudeau M. C. (eds.), *Handbook of Ion Channels*. Routledge Handbooks Online.

Frankfurt H. (1969). Alternative Possibilities and Moral Responsibility. *The Journal of Philosophy*, 66(23):829-839.
Frankfurt H. (1971). Freedom of the Will and the Concept of a Person. *The journal of Philosophy,* 68(1):5-20.
Frederick S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25-42.
Freedman S. J. & Clauser J. F. (1972). Experimental Test of Local Hidden-Variable Theory. *Physical Review Letters*, 28(14):938-941.
Fried I., Mukamel R. & Kreiman G. (2011). Internally Generated Preactivation of Single Neurons in Human Medial Frontal Cortex Predicts Volition. *Neuron,* 69:548-562.
Gerlich S., Eibenberger S., Tomandl M., Nimmrichter S., Hornberger K., Fagan P. J., Tüxen J., Mayor M. & Arndt M. (2011). Quantum Interference of Large Organic Molecules. *Nature Communications,* 2(1):263.
Gessell B. (2017). Indeterminism in the Brain. *Biology and Philosophy*, 32(6):1205-1223.
Gillies D. A. (1973). *An Objective Theory of Probability*. Routledge Revivals.
Ginet C. (1966). "Might we have no choice?" in K. Lehrer (ed.), *Freedom and Determinism*. Random House.
Ginet C. (1990). *On Action*. Cambridge University Press.
Ginet C. (1997). Freedom, Responsibility and Agency. *The Journal of Ethics,* 1:85-98.
Goldman A. (1992). "Epistemic folkways and scientific epistemology" in A. I. Goldman (ed.), *Liaisons: Philosophy Meets the Cognitive and Social Sciences*. MIT Press.
Goldman A. I. & Pust J. (1998). "Philosophical Theory and Intuitional Evidence" in Michael D. & William R. (eds.), *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*. Rowman & Littlefield.
Good I. J. (1983). "466656 Varieties of Bayesianism (#765)" in I. J. Good (ed.), *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press.
Greco J. (1992). Agent Reliabilism. *Noûs*, 33(13):273-296.
Greco J. (2003). Further Thoughts on Agent Reliabilism. *Philosophy and Phenomenological Research,* 66(2):466-480.
Gröblacher S., Paterek T., Kaltenbaek R., Brukner C., Zukowski M., Aspelmeyer M. & Zeilinger A. (2007). An Experimental Test of Non-local Realism. *Nature,* 446:871-875.
Hacking I. (1965). *Logic of Statistical Inference*. Cambridge University Press.
Haggard P. (2010). "Neuroscience, Free Will and Determinism: 'I'm Just a Machine'", Patrick Haggard interviewed by Tom Chivers for *The Telegraph,* 12. Oct. 2010. URL = https://www.telegraph.co.uk/news/science/8058541/Neuroscience-free-will-and-determinism-Im-just-a-machine.html
Hájek A. (2019). "Interpretations of Probability" *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). URL = https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/
Hanes D. P. & Schall J. D. (1996). Neural Control of Voluntary Movement Initiation. *Science,* 274(5286):427-430.
Hannikainen R., et al. (2019). For Whom Does Determinism Undermine Moral Responsibility? Surveying the Conditions for Free Will Across Cultures. *Frontiers in Psychology,* 10:2428.
Hare W. (1979). *Open-Mindedness and Education*. McGill-Queen's University Press.
Harris C. M. & Wolpert D. M. (1998). Signal-Dependent Noise Determines Motor Planning. *Nature,* 394(6695):780-784.
Hartmann S. & Sprenger J. (2010). "Bayesian Epistemology" in D. Pritchard & S. Bernecker (eds.), *The Routledge Companion to Epistemology*. Routledge.
Heatherton T. F. & Wagner D. D. (2011). Cognitive Neuroscience of Self-Regulation Failure. *Trends in Cognitive Science,* 15(3):132-139.
Heilbron J. L. (1986). *The Dilemmas of an Upright Man: Max Planck as Spokesman for German Science*. University of California Press.

Hertwig R. & Gigerenzer G. (1999). The 'Conjunction Fallacy' Revisited: How Intelligent Inferences Look Like Reasoning Errors. *Journal of Behavioral Decision Making*, 12(4):275-305.
Hesselmann G., Kell C. A., Eger E. & Kleinschmidt A. (2008). Spontaneous Local Variations in On-Going Neural Activity Bias Perceptual Decisions. *PNAS*, 105(31):10984-10989.
Hobbes T. (1651). *Leviathan*. Reprinted in R. E. Flatman & D. Johnston (eds.), Leviathan. W. W. Norton & Co.
Holton R. (1999). Intention and Weakness of Will. *Journal of Philosophy*, 96(5):241-262.
Hood B., Gjersoe N. L. & Bloom P. (2012). Do Children Think That Duplicating the Body also Duplicates the Mind? *Cognition*, 125(3):466-474.
Hook C. J. & Farah M. J. (2013). Look Again: Effects of Brain Images and Mind-Brain Dualism on Lay Evaluations of Research. *Journal of Cognitive Neuroscience*, 25(9):1397-1405.
Hornsby J. (2004). "Agency and Actions" in H. Stward and J. Hyman (eds.), *Agency and Actions*. Cambridge University Press.
Hume D. (1739). *A treatise of Human Nature*. Reprinted from the original edition in three volumes, by L. A. Selby-Bigge (eds.). Clarendon Press.
Ivanova M. (2010). Pierre Duhem's Good Sense as a Guide to Theory choice. *Studies in History and Philosophy of Science*, 41(1):58-64.
Jackson F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly*, 32(127):127-136.
Jackson F. (1986). What Mary Didn't Know. *The Journal of Philosophy*, 83(5):291-295.
Jacobson G. A., Diba K., Yaron-Jakoubovitch A., Oz Y., Koch C., Segev I. & Yarom Y. (2005). Subthreshold Voltage Noise of Rat Neocortical Pyramidal Neurons. *Journal of Physiology*, 564(pt 1):145-160.
James W. (1907). Pragmatism's Conception of Truth. *The Journal of Philosophy, Psychology and Scientific Methods*, 4(6):141-155.
Juffman T., Milic A., Müllneritsch M., Asenbaum P., Tsukernik A., Tüxen J., Mayor M., Chesnovsky O. & Arndt M. (2012). Real-time Single-molecule Imaging of Quantum Interference. *Nature Nanotechnology*, 7(5):297-300.
Kagaya K. & Takahata M. (2010). Readiness Discharge for Spontaneous Initiation of Walking in Crayfish. *Journal of Neuroscience*, 30(4):1348-1362.
Kane R. (1996). *The Significance of Free Will*. Oxford University Press.
Kane R. (1999). Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. *Journal of Philosophy*, 96(5)217–240.
Kane R. (2002). *The Oxford Handbook of Free Will*. Oxford University Press.
Kane R. (2007). Free Will: New Directions for an Ancient Problem: A Reply to Allen and Rogers. *Proceedings of the American Catholic Philosophical Association*,81:291-302.
Kastrup B. (2020). *Yes, Free Will Exists*. Scientific America. URL = https://blogs.scientificamerican.com/observations/yes-free-will-exists/
Kauppinen A. (2007). The Rise and Fall of Experimental Philosophy. *Philosophical Explorations* 10(2):95-118.
Kaye D. H. (1988). What is Bayesianism? A Guide for the Perplexed. *Jurimetrics*, vol. 28.
Keas M. N. (2018). Systematizing the Theoretical Virtues. *Synthese*, 195(6):2761-2793.
Kendall M. G. (1949). On the Reconciliation of Theories of Probability. *Biometrika*. Biometrika Trust. 36(1/2):101-116.
Kennett J. & Smith M. (1996). Frog and Toad Lose Control. *Analysis*, 56(2):63-73.
Kenny A. (1971) "The Homunculus Fallacy" in Greene M. (ed.), *Interpretations of Life and Mind: Essays around the problem of reduction*. Humanities Press.
Kidd J. (2011). Pierre Duhem's Epistemic Aims and the Intellectual Virtue of Humility: A Reply to Ivanova. *Studies in History and Philosophy of Science*, 42(1):185-189.
Kim J. (1992). "Downward Causation, Emergentism and Non-reductive Physicalism" in A. Beckermann, H. Flohr & J. Kim (Eds.), *Emergence or Reduction? Essays on the Prospect of Nonreductive Physicalism*. De Gruyter.

Kim J. (2005). *Physicalism or something Near Enough*. Princeton University Press.
Kim SI. (2013). Neuroscientific Model of Motivational Process. *Frontiers in Psychology*, 4:98.
Knobe J. (2007). Experimental Philosophy and Philosophical Significance. *Philosophical Explorations*, 10(2):119-121.
Knobe J. & Nichols S. (2017). "Experimental Philosophy", *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/
Kole M. H., Hallerman S. & Stuart G. J. (2006). Single Ih Channels in Pyramidal Neuron Dendrites: Properties, Distribution, and Impact on Action Potential output. *Journal of Neuroscience*, 26(6):1677-1687.
Kornhuber H. H. & Deecke L. (1965). Changes in Brain Potentials With Willful and Passive Movements in Humans: The Readiness Potential and Reafferent Potentials. *Pflugers Archieve*, 284:1-17.
Kurtz-David V., Persitz D., Webb R. & Levy D. J. (2019). The Neural Computation of Inconsistent Choice Behavior. *Nature Communications*, 10(1):1-14.
Kyberg H. E. (1974). Propensities and Probabilities. *The British Journal for the Philosophy of Science*, 25(4):358-375.
Lau M. YK. (2007). Extreme Response Style: An Empirical Investigation of the Effect of Scale Response Format and Fatigue. (Publication no. 3299156) [Doctoral Dissertation, University of Notre Dame], ProQuest Dissertations.
Lecoutre M., Rovira K., Lecoutre B. & Poitevineau J. (2005). People's Intuitions About Randomness and Probability: An Empirical Study. *Statistical Education Research Journal*, 5(1):1583.
Lewis J. R. (2019). Comparison of Four TAM Item Formats: Effects of Response Option Labels and Order. *Journal of Usability Studies*, 14(4):224-236.
Libet B., Gleason C. A., Wright E. W. & Pearl D. K. (1983). Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (readiness-potential). The Unconscious Initiation of a Freely Voluntary Act. *Brain*, 106(pt 3):623-642.
List C. (2014). Free Will, Determinism, and the Possibility of Doing Otherwise. *Nous*, 48(1):156-178.
List C. (2019). *Why Free Will is Real*. Harvard University Press.
Lowe E. J. (2008). *Personal Agency: The Metaphysics of Mind and Action*. Oxford University Press.
Loyola-Navarro R., Moenne-Loccoz C., Vergara R. C., Hyafil A. Aboitiz F. & Maldonado P. E. (2020). Agency Improves Working Memory and Accelerates Visual and Attentional Processing. *BioRxiv*.
Lyengar S. (1990). Framing Responsibility for Political Issues: The Case of Poverty. *Political Behaviour*, 12(1):19-40.
Manwani A. & Koch C. (1999). Detecting and Estimating Signals in Noisy Cable Structures. *Neural Computation*, 11(8):1831-1873.
McDonnell M. D. & Ward M. L. (2011). The Benefits of Noise in Neural Systems: Bridging Theory and Experiment. *Nature Reviews Neuroscience*, 12(7):415-426.
McFadden D. (2001). Economic Choices. *American Economic Review*, 91(3):351-378.
McKenna M. & Pereboom D. (2016). *Free Will: A Contemporary Introduction*. Routledge.
Mele A. (1992a). *Springs of Action*. Oxford University Press.
Mele A. (1992b). Acting for reasons and acting intentionally. *Pacific Philosophical Quarterly*, 73(4):355-374.
Mele A. (1997). Agency and Mental Action. *Philosophical perspectives*, 11:231-249.
Mele A. (1999). Kane, Luck, and the Significance of Free Will. *Philosophical Explorations*, 2(2):96-104
Mele A. (2001). "Acting Intentionally: Probing Folk Notions" in Malle B., Moses L. & Baldwin D. (eds.), *Intentions and Intentionality: Foundations of Social Cognition*. MIT Press.
Mele A. (2003). Motivation and Agency. Oxford University Press.
Mele A. (2006). *Free Will and Luck*. Oxford University Press.
Mele A. (2014). "Free Will and Substance Dualism: The Real Scientific Threat to Free Will?" in W. Sinnot-Armstrong (ed.), *Moral Psychology, Vol. 4: Free Will and Responsibility*, MIT Press.

Mele A. (2015). On Pereboom's Disappearing Agent Argument. *Criminal Law and Philosophy*, 11(3):561-574.
Mele A. & Moser P. K. (1994). Intentional Action. *Noüs*, 28(1):39-68.
Melden A. I. (1961). *Free Action*. Routledge and Kegan Paul.
Mellor D. H. (1971). *The Matter of Chance*. Cambridge University Press.
Mohan A., Pendyam S., Kalivas P. W. & Nair S. S. (2011). Molecular Diffusion Model of Neurotransmitter Homeostasis Around Synapses Supporting Gradients. *Neural Computation*, 23(4):984-1014.
Monroe A. & Malle B. (2010). From Uncaused Will to Conscious Choice: The Need to Study, Not Speculate about People's Folk Concept of Free Will. *Review of Philosophy and Psychology*, 1(2):211-224.
Monroe A., Dillon K. & Malle B. (2014). Bringing Free Will Down to Earth: People's Psychological Concept of Free Will and Its Role in Moral Judgment. *Consciousness and Cognition*, 27c(1):100-108.
Montague P. R. (2008). Free Will. *Current biology*, 18(14):R584-R585.
Monterosso J., Boyzman E. & Schwartz B. (2005). Explaining Away Responsibility: Effects of Scientific Explanation on Perceived Culpability. *Ethics & Behavior*, 15(2):139-158.
Moors G., Kieruj N. D. & Vermut J. K. (2014). The Effect of Labeling and Numbering of Response Scales on the Likelihood of Response Bias. *Sociological Methodology*, 44(1):369-399.
Murray D. & Nahmias E. (2012). Explaining Away Incompatibilist Intuitions. *Philosophy of Phenomenological Research*, 88(2):434-467.
Nadelhoffer T. & Nahmias E. (2007). The Past and Future of Experimental Philosophy. *Philosophical Explorations*, 10(2):123-149.
Nadelhoffer T., Rose D., Buckwalter W. & Nichols S. (2020). Natural Compatibilism, Indeterminism, and Intrusive Metaphysics. *Cognitive Science*, 44(8):e12873.
Nadelhoffer T., Shepard J., Nahmias E., Sripada C. & Ross L. T. (2014). The Free Will Inventory: Measuring Beliefs about Agency and Responsibility. *Consciousness and Cognition*, 25(1):27-41.
Nadelhoffer T., Rose D., Buckwalter W. & Nichols S. (2020). Natural Compatibilism, Indeterminism, and Intrusive Metaphysics. *Cognitive Science*, 44(8):E12873.
Nagel T. (1986). *The View from Nowhere*. Oxford University Press.
Nahmias E. (2006). Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism. *Journal of Cognition and Culture*, 6(1):215-237.
Nahmias E., Coates D. & Kvaran T. (2007). Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions. *Midwest Studies in Philosophy*, 31(1):214-242.
Nahmias E., Morris S., Nadelhoffer T. & Turner J. (2005). Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility. *Philosophical Psychology*, 18(5):561-584.
Nahmias E. & Murray D. (2010). "Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions" in Aguilar J., Buckareff A. & Frankish K. (eds.), *New Waves in Philosophy of Action*. Palgrave-Macmillan.
Nahmias E., Shepard J. & Reuter S. (2014). It's OK if 'my brain made me do it': People's Intuitions about Free Will and Neuroscientific Prediction. *Cognition*, 133(2):502-516.
Nairz O., Arndt M. & Zeilinger A. (2003). Quantum Interference Experiments with Large Molecules. *American Journal of Physics*, 71(4):319-325.
Neurath O. (1931). "Physicalism: The Philosophy of the Vienna Circle" in R. S. Cohen & M. Neurath (eds.), *Philosophical Papers 1913–1946*. D. Reidel Publishing Company
Neuringer A. (1986). Can People Behave "Randomly?" The Role of Feedback. *Journal of Experimental Psychology: General*, 115(1):62-75.
Nichols S. (2004). The Folk Psychology of Free Will: Fits and Starts. *Mind and Language*, 19(5):473-502.
Nichols S. (2006a). Folk Intuitions on Free Will. *Journal of Cognition and Culture*, 6(1-2):57-86.

Nichols S. (2006b). Free Will and the Folk: Responses to Commentators. *Journal of Cognition and Culture*, 6(1):305-320.
Nichols S. & Knobe J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Nous*, 41(4):663-685.
Nichols S. & Rose D. (2013). The Lesson of Bypassing. *Review of Philosophy and Psychology*, 4(4):599-619.
Norenzayan A. & Lee A. (2010). It Was Meant to Happen: Explaining Cultural Variations in Fate Attributions. *Journal of Personality and Social Psychology*, 98(5):702-20.
O'Connor T. (2000). *Persons and Causes*. Oxford University Press.
O'Connor T. (2007). Is it All Just a Matter of Luck? *Philosophical Explorations*, 10(2):157-161.
O'Connor T. & Christopher F. (2020). "Free Will", *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.). URL = https://plato.stanford.edu/archives/spr2020/entries/freewill/
Paterek T., Fedrizzi A., Gröblaucher S., Jennewein T., Zukowski M., Aspelmeyes M. & Zellinger A. (2007). An Experimental Test of Non-local Realistic Theories without the Rotational Symmetry Assumption. *Physical Review Letters*, 99(21):210406.
Paulhus D. & Carey J. (2011). The FAD-Plus: Measuring Lay Beliefs Regarding Free Will and Related Constructs. *Journal of Personality Assessment*, 93(1):96-104.
Pereboom D. (2001). *Living Without Free Will*. Cambridge University Press.
Pereboom D. (2013). "Free Will Skepticism and Criminal Punishment", in Thomas Nadelhoffer (ed.), *The Future of Punishment*. Oxford University Press.
Pereboom D. (2014a). *Free Will, Agency, and Meaning in Life*. Oxford University Press.
Pereboom D. (2014b). Is our Conception of Agent-Causation Coherent? *Philosophical Topics*, 32(1/2):275-286.
Pereboom, D. (2017). "Responsibility, Agency, and the Disappearing Agent Objection" in JB. Guilon (ed.), *Le libre arbiter*. Collége de France.
Pereboom D. & Caruso G. (2002). "Hard Incompatibilist Existentialism: Neuroscience, Punishment, and Meaning in Life" in Caruso D. & Flanagan O. (eds.), *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*. Oxford University Press.
Pierce C. S. (1931). *The Collected Papers of Charles Sanders Pierce, Vol. 1: The Principles of Philosophy*. Harvard University Press.
Pink T. (2004). *Free Will: A Very Short Introduction*. Oxford University Press.
Planck M. (1958). *Physikalische Abhandlungen und Vorträge*. Vieweg.
Platkiewicz J. & Brette R. (2010). A Threshold Equation for Action Potential Initiation. *PLoS Computational Biology*, 6(7):e1000850.
Popper K. R. (1959). The Propensity Interpretation of Probability. *The British Journal for the Philosophy of Science*, 10(37):25-42.
Pothos E. M., Busemeyer J. R. & Trueblood J. S. (2013). A Quantum Geometric Model of Similarity. *Psychological Review*, 120(3):679-696.
Preston L., Ritter S. & Hepler J. (2013). Neuroscience and the Soul: Competing Explanations for the Human Experience. *Cognition*, 127(1):31-37.
Pust J. (2019). Intuition, *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.) URL = https://plato.stanford.edu/archives/sum2019/entries/intuition/
Quine W. V. O. (1951). Two Dogmas of Empiricism. *The Philosophical Review*, 60(1):20-43.
Quine W. V. O. (1955). "Posits and Reality" in *The Ways of the Paradox and Other Essays*. Harvard University Press.
Quine W. V. O. (1981). *Theories and Things*. Harvard University Press.
Ravizza M. (1994). Semi-Compatibilism and the Transfer of Non-Responsibilit. *Philosophical Studies*, 75(1-2):61-93.
Rawls J. (1971). *A Theory of Justice*. Harvard University Press.
Rieke F. & Baylor D. (1998). Origin of Reproducibility in the Response of Retinal Rods to Single Photons. *Biophysical Journal*, 75(4):1836-1857.

Rieke F., Warland D., de Reuter van Steveninck R. & Bialek W. (1997). *Spikes: Exploring the Neural Code*. MIT Press.
Riggs W. (2010). Open-Mindedness. *Metaphilosophy*, 41(1-2):172-188.
Robinson W. (2019). "Epiphenomenalism", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). URL = https://plato.stanford.edu/archives/sum2019/entries/epiphenomenalism/
Romo R. & Schultz W. (1987). Neuronal Activity Preceding Self-Initiated or Externally Timed Arm Movements in Area 6 of Monkey Cortex. *Experimental Brain Research*, 67:656-662.
Rose D., Bukwaiter W., & Nichols S. (2015). Neuroscientific Prediction and the Intrusion of Intuitive Metaphysics. *Cognitive Science*, 41(2):482-502.
Roskies A. (2006). Neuroscientific Challenges to Free Will and Responsibility. *Trends in Cognitive Sciences*, 10(9):419-423.
Roskies A. (2010). How Does Neuroscience Affect Our Conception of Volition? *Annual Review of Neuroscience*, 33(1):109-130.
Ryle G. (1949). *The Concept of Mind*. University of Chicago Press.
Saarinen A., Linne M. L. & Yli-Harja O. (2008). Stochastic Differential Equation Model for Cerebellar Granule Cell Excitability. *PLOS Computational Biology*, 4(2):e1000004.
Sarkissian H., Chatter Jee A., De Brigard F., Knobe J., Nichols S. & Sirker S. (2010). Is Belief in Free Will a Cultural Universal? *Cognition*, 25(3):346-358.
Schilpp P. A. (1998). *Albert Einstein, Philosopher-scientist*. Mjf books.
Schlosser M. E. (2007). Basic Deviance Reconsidered. *Analysis*, 67(3):183-194.
Schlosser M. E. (2010). "Agency, Ownership and the Standard Theory" in J. Aguilar, A. Buckareff & K. Frankish (eds.), *New Waves in Philosophy of Action*. Palgrave-Macmillian.
Schneidman E., Freedman B. & Segev I. (1998). Ion Channel Stochasticity may be Critical in Determining the Reliability and Precision of Spike Timing. *Neural Computation and Applications*, 10(7):16791703.
Schopenhauer A. (1839). *On The Freedom of the Will*. Reprinted as "Prize Essay on the Freedom of the Will", translation by Eric F. J. Payne. Cambridge University Press.
Schrödinger E. (1935). Die Gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften*, 23(50):844-849.
Schurger A., Sitt J. D. & Dehaene S. (2012). An Accumulator Model for Spontaneous Neural Activity Prior to Self-initiated Movement. *PNAS*, 109(42):E2904-E2913.
Seamus B. (2019). "Imprecise Probabilities", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2019/entries/imprecise-probabilities/
Shadlen M. N. & Newsome W. T. (1998). The Variable Discharge of Cortical Neurons: Implications for Connectivity, Computation, and Information Coding. *The Journal of Neuroscience*, 18(10):3870-3896.
Shadlen M. N. & Roskies A. L. (2012). The Neurobiology of Decision-Making and Responsibility: Reconciling Mechanism and Mindedness. *Frontiers in Neuroscience*, 6:56.
Shapiro L. (1999). Princess Elizabeth and Descartes: The Union of Soul and Body and the Practice of Philosophy. *British Journal for the History of Philosophy*, 7(3):503-520.
Shultz E., Cokely E. & Feltz A. (2011). Persistent Bias in Expert Judgments about Free Will and Moral Responsibility: A Test of the Expertise Defense. *Consciousness and Cognition*, 20(4):1722-1731.
Smilansky S. (2000). *Free Will and Illusion*. Oxford University Press.
Sommers T. (2009). More Work For Hard Incompatibilism. *Philosophy and Phenomenological Research*, LXXIX.3:511-521.
Sommers T. (2010). Experimental Philosophy and Free Will. *Philosophy Compass* 5(2):199-212.
Sosa E. (1980). The Raft and the Pyramid: Coherence Versus Foundations in the Theory of Knowledge. *Midwest Studies in Philosophy*, 5(1):3–26.
Spinoza B. (1677). *Ethics*. Penguin Classics.
Sripada C. S. (2012). How is Willpower Possible? The Puzzle of Synchronic Self-Control and the Divided Mind. *Noûs*, 48(1):41-74.

Stanford K. (2017). "Underdetermination of Scientific Theory", *The Standford Encyclopedia of Philosophy*. URL = https://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian/
Steward H. (2012). *A Metaphysics for Freedom*. Oxford University Press.
Steward H. (2013). Processes, Continuants and Individuals. *Mind*, 122(487):781-812.
Stoljar D. (2017). "Physicalism", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). URL = https://plato.stanford.edu/archives/win2017/entries/physicalism/
Strawson G. (1986). *Freedom and Belief*. Oxford University press.
Stump D. (2007). Pierre Duhem's Virtue Epistemology. *Studies in History and Philosophy of Science*, 38(1):149-159.
Südhof C. T. (2013). Neurotransmitter Release: The Last Milisecond in the Life of a Synaptic Vesicle. *Neuron*, 80(3):675-690.
Suppes P. (1993). The Transcendental Character of Determinism. *Midwest Studies in Philosophy*, 18(1):242–257.
Talbott W. (2016). "Bayesian Epistemology", *The Standford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). URL = https://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian/
Taylor R. (1966). *Metaphysics*. Prentice-hall inc.
Taylor S. D. (2019). Two Kinds of Explanatory Integration in Cognitive Science. *Synthese*.
Tolhurst D. J., Movshon J. A. & Dean A. F. (1981). The Statistical Reliability of Signals in Single Neurons in Cat and Monkey Visual Cortex. *Vision Research*, 23(8):775-785.
Tse P. (2011). *The Neural Basis of Free Will*. MIT Press.
Tsodyks M. V. & Markram H. (1997). The Neural Code Between Neocortical Pyramidal Neurons Depends on Neurotransmitter Release Probability. *PNAS*, 94(2):719-723.
Uddin Q. L. (2020). Bring the Noise: Reconceptualizing Spontaneous Neural Activity. *Trends in Cognitive Neuroscience*, 24(9):734-746.
Van Inwagen P. (1975). The Incompatibility of Free Will and Determinism. *Philosophical Studies*, 27(3):185-199.
Van Inwagen P. (1983). *An Essay on Free Will*. Oxford University Press.
Van Inwagen P. (2017). *Thinking About Free Will*. Cambridge University Press.
Varela A. J., Sen K., Gibson J., Fost J., Abbott L. F. & Nelson B. S. (1997). A Quantitative Description of Short-Term Plasticity at Excitatory Synapses in Layer 2/3 of Rat Primary Visual Cortex. *Journal of Neuroscience*, 17(20):7926-7940.
Velleman D. (1992). What Happens When Someone Acts? *Mind*, 101(403):461-481.
Vonasch A. J., Baumeister R. F. & Mele, A. (2018). Ordinary People Think Free Will is a Lack of Constraint, not the Presence of a Soul. *Consciousness and Cognition*, 60:133-151.
Waller B. N. (1988) Free Will Gone Out of Control. *Behaviourism*, 16(2):149-167.
Wang L., Wang H., Yu L. & Chen Y. (2016). Spike-threshold Variability Originated from Separatrix-Crossing in Neuronal Dynamics. *Scientific reports*, 6:31719.
Weber M. (2005). Indeterminism in Neurobiology. *Philosophy of Science*, 72(5):663-674.
Wegner D. M. (2002). *The Illusion of Conscious Will*. MIT Press.
Weijters B., Cabooter E. & Schillewaert N. (2010). The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels. *International Journal of Research in Marketing*, 27(3):236-247.
Weiskopf D. A. (2017). "The Explanatory Autonomy of Cognitive Models" In D. M. Kaplan (ed.), *Explanation and Integration in Mind and Brain Science*. Oxford University Press.
Werndl C. (2009). Are deterministic Descriptions and Indeterministic Descriptions Observationally Equivalent? *Studies in History and Philosophy of Modern Physics*, 40(3):232-242.
Werndl C. (2011). On the Observational Equivalence of Continuous-time Deterministic and Indeterministic Descriptions. *European Journal for the Philosophy of Science*, 1(2):193-225.
Werndl C. (2016). "Determinism and Indeterminism" in P. Humphreys (ed), *The Oxford Handbook of Philosophy of Science*. Oxford University Press.

Wilson D. L. (1999). Mind-Brain Interaction and Violation of Physical Laws. *Journal of Consciousness Studies*, 6(8-9):185-200.
Wisniewski D., Deutschländer R. & Haynes J-D. (2019). Free Will Beliefs are Better Predicted by Dualism than Determinism Beliefs Across Different Cultures. *PLoS ONE*, 14(9):e0221617.
Wittgenstein L. (1969). *On Certainty*. Basil Blackwell.
Wolf S. (1990). *Freedom Within Reason*. Oxford University Press.
Zagzebski L. (1996). *Virtues of the Mind*. Cambridge University Press.

Curriculum Vitae

Academic Degrees

Current: Ph.D. in Neurophilosophy, Graduate School of Systemic Neuroscience at LMU Munich. Primary Supervisor: Ophelia Deroy. Co-supervisors: Stephan Sellmaier & Paul C. J. Taylor

2017 Master's degree in philosophy, University of Copenhagen
Thesis: "*Between Intention and Action: A Problem of Causal Pathways"*

2015 Bachelor's degree in philosophy, University of Copenhagen

Academic Focus

Epistemology, Philosophy of Action, Philosophy of Science, Philosophy of Mind, Philosophy of Language, Cognitive science, Neuroscience, Experimental Philosophy, Psychology

Relevant Employment

2018 Intern, Dept. of Psychology, University of Copenhagen

2017 Research Assistant, 'Intentions, Selection, and Agency', Dept. of Philosophy, University of Copenhagen

Teaching experience:

2021 "Debating Club Seminar" (Course on social neuroscience), NCP/CVBE/LMU Munich

2020 "Responsibility in mind: Free will, control and moral agency" (Course on experimental philosophy and free will), CVBE/LMU Munich

2020 "Knowing that we know: Key issues on reflection and confidence" (Course on Metacognition), CVBE/LMU Munich

Conferences, talks and posters

2019 Summer School: Experimental Methods in Philosophy, Venice, Italy

2019 Summer School: Recent Developments in Situated Cognition, Bochum, Germany

2019 Workshop: Models of Control and Moral Responsibility, Copenhagen, Denmark

2019: Summer School: Norms & Biases in Social Interactions, Aegina, Greece

2019: Symposium: Biology of Decision-making, Oxford, UK

2018: Summer School: New Perspectives & Methods on Social Cognition, Aegina, Greece

2018: Summer School: The Perception/Cognition Divide, Venice, Italy

2018: Workshop: Mechanisms and Models of Intention, Cambridge, UK

2017: Workshop: Mechanisms of Intention, Copenhagen, Denmark

2017: Workshop: Agency and Control, Copenhagen, Denmark

Technical skills

Electroencephalogram (EEG)

R-studio

JavaScript

Languages

Danish (Native)

English (Fluent)

German (B1)

List of Publications

Mark Wulff Carstensen. (Under Review). **Epistemic Virtues and Underdetermination**. *Noús.*

Mark Wulff Carstensen, Ophelia Deroy, Stephan Sellmaier, Paul C. J. Taylor (Finished Manuscript). **Free Will Properly Understood: Folk-Conceptions are Incompatibilist, but not Hard Incompatibilist.**

Mar Wulff Carstensen, Ophelia Deroy, Stephan Sellmaier, Paul C. J. Taylor (Finished Manuscript). **Belief in physicalism or Belief in Substance-Dualism, Free Will is Incompatibilist.**

Mark Wulff Carstensen (Finished Manuscript). **Recursive Decisions, Indeterminism and Luck**.

Affidavit/Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation *Free Will and Chance: Towards a Graded Approach to Agency* selbständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation *Free Will and Chance: Towards a Graded Approach to Agency* is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

27.09.2021
München, den
Munich, date

Mark Wulff Carstensen
Unterschrift
Signature

Declaration of Author Contributions

I declare that this thesis has been composed solely by myself, that the work contained herein is the result of my own work, and that this work has not been submitted for any other degree of professional qualification.

The contents of chapter 2 *Free Will Properly Understood: Folk-Conceptions are Incompatibilist, but not Hard Incompatibilist* are part of a manuscript for which I am the first author.

The contents of chapter 3 *Belief in Physicalism or Belief in Substance-Dualism: Free Will is Incompatibilist* are part of a manuscript for which I am the first author.

The contents of chapter 4 *Epistemic Virtues and Underdetermination: When Should We Be Indeterminists?* Are part of a manuscript for which I am the first and sole author.

The contents of chapter 5 *Recursive Decisions, Indeterminism, and Luck* are part of a manuscript for which I am the first and sole author.

Munich, May 2021

Mark Wulff Carstensen

Prof. Dr. Ophelia Deroy